Contraction of the proximal map and generalized convexity of the Moreau-Yosida regularization in the 2-Wasserstein metric

Eric A. Carlen[1] and Katy Craig[2]

Department of Mathematics, Hill Center,
Rutgers University, 110 Frelinghuysen Road Piscataway NJ 08854-8019 USA

October 10, 2012

Abstract

We investigate the Moreau-Yosida regularization and the associated proximal map in the context of discrete gradient flow for the 2-Wasserstein metric. Our main results are a stepwise contraction property for the proximal map and an "above the tangent line" inequality for the regularization. Using the latter, we prove a Talagrand inequality and an HWI inequality for the regularization, under appropriate hypotheses. In the final section, the results are applied to study the discrete gradient flow for Rényi entropies. As Otto showed, the gradient flow for these entropies in the 2-Wasserstein metric is a porous medium flow or a fast diffusion flow, depending on the exponent of the entropy. We show that a striking number of the remarkable features of the porous medium and fast diffusion flows are present in the discrete gradient flow and do not simply emerge in the limit as the time-step goes to zero.

Key Words: Wasserstein metric, gradient flow, Moreau-Yosida regularization.

1 Introduction

Given a complete metric space (X, d), a functional $E : X \to \mathbb{R} \cup \{\infty\}$, and $\tau > 0$, the *Moreau-Yosida regularization* of E is

$$E_\tau(y) := \inf_{x \in X} \left\{ \frac{1}{2\tau} d(x, y)^2 + E(x) \right\}.$$

The corresponding *proximal set* $J_\tau : X \to 2^X$ is

$$J_\tau(y) := \operatorname*{argmin}_{x \in X} \left\{ \frac{1}{2\tau} d(x, y)^2 + E(x) \right\}.$$

If there is a unique element in $J_\tau(y)$, we denote it by y_τ and call it the *proximal point*. We call $y \mapsto y_\tau$ the *proximal map*.

[1]Work partially supported by U.S. National Science Foundation grant DMS 0901632.
[2]Work partially supported by a Presidential Fellowship at Rutgers University

When $X = \mathcal{H}$ is a Hilbert space, a suitable context in which to develop the theory of the Moreau-Yosida regularization is the class of functionals that are proper, lower semicontinuous, and convex. For all such E and $\tau > 0$, the Moreau-Yosida regularization E_τ is convex and Fréchet differentiable [16]. Furthermore, its derivative is Lipschitz continuous, and, as $\tau \to 0$, $E_\tau \nearrow E$ pointwise [6]. The Moreau-Yosida regularization provides a way to regularize E that preserves convexity.

The proximal map is similarly well-behaved for functionals that are proper, lower semicontinuous, and convex. For each $y \in \mathcal{H}$ and $\tau > 0$, there is a unique proximal point y_τ, so that the proximal map $y \mapsto y_\tau$ is well-defined on all of $\mathcal{H}$. As shown by Moreau [16], the proximal map is a contraction in the Hilbert space norm:

$$||x_\tau - y_\tau|| \leq ||x - y|| \ \forall x, y \in \mathcal{H}.$$

One of the main reasons for interest in the Moreau-Yosida regularization and proximal map is their relation to gradient flow. The *gradient flow* of a functional E is the Cauchy problem

$$\frac{d}{dt}y(t) = -\nabla E(y(t)), \quad y(0) \in \overline{D(E)} = \overline{\{z \in \mathcal{H} : E(z) < \infty\}}, \tag{1.1}$$

which is well-defined as long as ∇E exists along the flow $y(t)$.[1] The Moreau-Yosida regularization plays a key role in the proof of existence for solutions to the gradient flow [5]. First, one uses the additional regularity of E_τ to find solutions to the related gradient flow problem

$$\frac{d}{dt}y_\tau(t) = -\nabla E_\tau(y_\tau(t)), \quad y_\tau(0) \in \overline{D(E)}.$$

Then, as $\tau \to 0$, the curves $y_\tau(t)$ converge to a curve $y(t)$ that solves (1.1) in an appropriate sense.

The proximal map expresses the discrete dynamics of gradient flow. Specifically, one may use the proximal map to define the *discrete gradient flow* sequence

$$y_n = (y_{n-1})_\tau, \quad y_0 \in \overline{D(E)},$$

as in [12, 13]. Whenever the proximal map $y \mapsto y_\tau$ is well-defined, we may identify the proximal set $J_\tau(y)$ with its unique element y_τ and write J_τ^n to indicate n repeated applications of the proximal map. The exponential formula quantifies the sense in which the discrete gradient flow is a discretized version of gradient flow [6]. If $y(t)$ is a gradient flow with initial conditions $y(0)$, then

$$y(t) = \lim_{n\to\infty} (J_{t/n})^n(y(0)). \tag{1.2}$$

More recently, the Moreau-Yosida regularization and proximal map have been applied outside of the Hilbert space context to gradient flow in the 2-Wasserstein metric. Briefly, we recall some facts about this metric, mainly to establish our notation — see [2] and [22] for more background. We present these facts both in the most general setting, without restrictions on the type of probability measures we consider, and in a simpler setting, focusing our attention on probability measures with finite second moment that are absolutely continuous with respect to Lebesgue measure. While our results hold in the most general setting, many interesting applications concern only the simpler setting, in which the exposition and notation is more straightforward.

[1] Alternatively, one may define the gradient flow in terms of the subdifferential [5].

Let $\mathcal{P}(\mathbb{R}^d)$ denote the set of Borel probability measures on $\mathbb{R}^d$. Given $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, a Borel map $T : \mathbb{R}^d \to \mathbb{R}^d$ *transports* μ *onto* ν if $\nu(B) = \mu(T^{-1}(B))$ for all Borel sets $B \subseteq \mathbb{R}^d$. We call ν the *push-forward of* μ *under* T and write $\nu = T\#\mu$.

Now consider a measure $\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$. (We will distinguish probability measures on $\mathbb{R}^d \times \mathbb{R}^d$, from probability measures on $\mathbb{R}^d$ by writing them in bold font.) Let π_1 be the projection onto the first component of $\mathbb{R}^d \times \mathbb{R}^d$, and let π_2 be the projection onto the second component. The first and second *marginals* of $\boldsymbol{\mu}$ are $\pi_1\#\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R}^d)$ and $\pi_2\#\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R}^d)$.

Given $\mu, \nu \in P(\mathbb{R}^d)$, the set of *transport plans* from μ to ν is

$$\Gamma(\mu, \nu) := \{\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : \pi_1\#\boldsymbol{\mu} = \mu \ , \pi_2\#\boldsymbol{\mu} = \nu\} \ .$$

The *2-Wasserstein distance* between μ and ν is

$$W_2(\mu, \nu) := \left(\inf \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 d\boldsymbol{\mu}(x, y) : \boldsymbol{\mu} \in \Gamma(\mu, \nu) \right\} \right)^{1/2} . \tag{1.3}$$

When $W_2(\mu, \nu) < \infty$, this infimum is attained, and we refer to the plans that attain the infimum as *optimal transport plans*. We denote the set of optimal transport plans by $\Gamma_0(\mu, \nu)$.

The 2-Wasserstein distance satisfies the triangle inequality and is non-negative, non-degenerate, and symmetric. However, $\mathcal{P}(\mathbb{R}^d)$ endowed with the 2-Wasserstein distance is not a metric space, since there exist measures that are infinite distances apart. Let $\mathcal{P}_{\mu_0}(\mathbb{R}^d)$ be the subset of $\mathcal{P}(\mathbb{R}^d)$ consisting of measures that are a finite distance from some fixed Borel probability measure μ_0, so that, by the triangle inequality, $(\mathcal{P}_{\mu_0}(\mathbb{R}^d), W_2)$ is a metric space. As indicated by the notation, one may take μ_0 to be the initial conditions of a gradient flow. Note that when $\mu_0 = \delta_0$, the Dirac mass at the origin, $\mathcal{P}_{\delta_0}(\mathbb{R}^d)$ is the subset of $\mathcal{P}(\mathbb{R}^d)$ with finite second moment.

We now define the 2-Wasserstein distance in a simpler setting. Let $P_2(\mathbb{R}^d)$ denote the set of probability measures with finite second moment and $\mathcal{P}_2^a(\mathbb{R}^d)$ denote the set of probability measures with finite second moment that are absolutely continuous with respect to Lebesgue measure. If $\mu \in P_2^a(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, the 2-Wasserstein distance between μ and ν reduces to the form

$$W_2(\mu, \nu) := \left(\inf \left\{ \int |x - T(x)|^2 d\mu(x) : T\#\mu = \nu \right\} \right)^{1/2} . \tag{1.4}$$

The Brenier-McCann theorem guarantees that the infimum in (1.4) is attained by $T = \nabla\varphi$, where $\varphi : \mathbb{R}^d \to \mathbb{R}$ is convex and $\nabla\varphi$ is unique μ-almost everywhere [15]. In particular,

$$W_2^2(\mu, \nu) = \int |x - \nabla\varphi(x)|^2 d\mu(x) \ ,$$

and we call $\nabla\varphi$ the *optimal transport map from* μ *to* ν. To emphasize its dependence on μ and ν, we denote the optimal transport map from μ to ν by $\mathbf{t}_\mu^\nu$.

Given $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^d)$ with $W_2^2(\mu_1, \mu_2) < \infty$ and $\boldsymbol{\mu} \in \Gamma_0(\mu^1, \mu^2)$, a *geodesic* connecting μ_1 and $\mu_2 \in \mathcal{P}(\mathbb{R}^d)$ is a curve of the form

$$\mu_\alpha^{1\to 2} : [0, 1] \to \mathcal{P}(\mathbb{R}^d), \quad \mu_\alpha^{1\to 2} = ((1 - \alpha)\pi_1 + \alpha\pi_2)\,\#\boldsymbol{\mu} \ .$$

As shown in [2, Theorem 7.2.2], this definition agrees with the metric space definition of a geodesic, i.e. a curve $\mu_\alpha : [0, 1] \to \mathcal{P}(\mathbb{R}^d)$ with $W_2(\mu_0, \mu_1) < \infty$ such that $W_2(\mu_\alpha, \mu_\beta) = |\alpha - \beta| W_2(\mu_0, \mu_1)$. If $\mu_1 \in \mathcal{P}_2^a(\mathbb{R}^d)$, $\mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, then the geodesic connecting μ_1 and μ_2 is unique and of the form

$$\mu_\alpha^{1\to 2} : [0, 1] \to \mathcal{P}_2(\mathbb{R}^d), \quad \mu_\alpha^{1\to 2} = \left((1 - \alpha)\mathrm{id} + \alpha\mathbf{t}_{\mu_1}^{\mu_2}\right)\#\mu_1,$$

where $\mathrm{id}(x) = x$ is the identity transformation.

A functional $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ is λ-*convex* in the 2-Wasserstein metric if, for all $\mu_1, \mu_2 \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)$, there exists a geodesic connecting μ_1 and μ_2 along which E is λ-convex:

$$E(\mu_\alpha^{1\to 2}) \leq (1-\alpha)E(\mu_1) + \alpha E(\mu_2) - \alpha(1-\alpha)\frac{\lambda}{2}W_2^2(\mu_1,\mu_2). \tag{1.5}$$

If a functional is 0-convex, we simply call it *convex*.[2] If a functional is 0-convex and strict inequality holds in (1.5) for all $\alpha \in (0,1)$, we call it *strictly convex*.

Given a functional $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ and $\tau > 0$, its Moreau-Yosida regularization is

$$E_\tau(\mu) := \inf_{\nu \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)} \left\{ \frac{1}{2\tau}W_2^2(\mu,\nu) + E(\nu) \right\} \tag{1.6}$$

and the corresponding *proximal set* $J_\tau : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to 2^{\mathcal{P}_{\mu_0}(\mathbb{R}^d)}$ is

$$J_\tau(\mu) := \operatorname*{argmin}_{\nu \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)} \left\{ \frac{1}{2\tau}W_2^2(\mu,\nu) + E(\nu) \right\} . \tag{1.7}$$

As before, if there is a unique element in $J_\tau(\mu)$, we denote it by μ_τ and call it the *proximal point*. Similarly, we call $\mu \mapsto \mu_\tau$ the *proximal map*. The properties of the Moreau-Yosida regularization and proximal map in the 2-Wasserstein metric will be the main focus of this paper.

As in the Hilbertian case, one of the main reasons for interest in the Moreau-Yosida regularization and the proximal map in the 2-Wasserstein metric is their relation to gradient flow. When E and μ are sufficiently smooth, the *2-Wasserstein gradient of E at $\mu \in D(E)$* is

$$\nabla_W E(\mu) = -\nabla \cdot \left(\mu \nabla \frac{\delta E}{\delta \rho}(\mu) \right), \tag{1.8}$$

where $\frac{\delta E}{\delta \rho}$ is the functional derivative of E [19] [2, Chapters 8 and 10].[3] The *gradient flow* of E is the Cauchy problem

$$\frac{d}{dt}\mu(t) = -\nabla_W E(\mu(t)), \quad \mu(0) \in \overline{D(E)} = \overline{\{\mu \in P_{\mu_0}(\mathbb{R}^d) : E(\mu) < \infty\}},$$

which is well-defined, as long as $\nabla_W E(\mu(t))$ exists along the flow $\mu(t)$[4]. We will sometimes refer to this as the *continuous gradient flow* in order to distinguish it from the *discrete gradient flow* we define below.

Otto observed that $-\nabla \cdot \left(\mu \nabla \frac{\delta E}{\delta \rho}(\mu) \right)$ may be viewed as the gradient vector field on the "Riemannian manifold of probability densities on $\mathbb{R}^d$" associated to the functional E, where the Riemannian metric is the infinitesimal form of the 2-Wasserstein metric [18, 19]. (It is one of his insights that the 2-Wasserstein metric is induced by a Riemannian metric.) In this metric, the length of the gradient of E at μ is given by

$$|\nabla_W E(\mu)| = \left(\int \left| \nabla \frac{\delta E}{\delta \rho}(\mu) \right|^2 d\mu \right)^{1/2} . \tag{1.9}$$

[2] It is also common to refer to convex functionals in the 2-Wasserstein metric as *displacement convex* [14].

[3] Some authors – e.g. [2] – identify the tangent vector $\nabla_W E(\mu)$ with the gradient vector field $-\nabla \frac{\delta E}{\delta \rho}(\mu)$ on $\mathbb{R}^d$. One gets Otto's representative from this by multiplying by μ and taking the divergence. The choice of representatives is merely notational.

[4] Alternatively, one may define gradient flow in terms of the subdifferential [2, Definition 11.1.1].

As in the Hilbertian case, the proximal map expresses the dynamics for discrete gradient flow. When the proximal map $\mu \mapsto \mu_\tau$ is well-defined (which occurs under much weaker assumptions on E and μ than are needed to define the gradient, as we describe before equation (1.14) below) we may define the *discrete gradient flow* sequence

$$\mu_n = (\mu_{n-1})_\tau, \quad \mu_0 \in \overline{D(E)} \ . \tag{1.10}$$

As before, we identify the proximal set $J_\tau(\mu)$ with its unique element μ_τ and write J_τ^n to indicate n repeated applications of the proximal map.

One of the advantages of discrete gradient flow is that it is not necessary to make precise the sense in which (1.8) defines a gradient vector field. This fact was emphasized by De Giorgi in his theory of the metric derivative [8] and extensively developed by Ambrosio, Gigli, and Savaré [2, Chapter 8]. We follow De Giorgi's lead, and all of the estimates we use involve only the length of the gradient $|\nabla_W E(\mu)|$. In the case that E and μ lack sufficient smoothness for (1.9) to be well-defined, we will interpret the symbol $|\nabla_W E(\mu)|$ as the *metric slope*

$$\limsup_{\nu \to \mu} \frac{(E(\mu) - E(\nu))^+}{W_2(\mu, \nu)} \ . \tag{1.11}$$

We use the heuristic notation $|\nabla_W E(\mu)|$ since, as demonstrated by Otto [18, 19], it is often enlightening to think of $|\nabla_W E(\mu)|$ as coming from a Riemannian metric on $\mathcal{P}(\mathbb{R}^d)$.

In their recent book [2], Ambrosio, Gigli, and Savaré conduct a detailed study of gradient flow and discrete gradient flow in the 2-Wasserstein metric for large classes of functionals, developing the analogy with the Hilbert space theory. It would be too much to hope for a perfect analogy. For example, in the Hilbert space context, if a functional E is proper, lower semicontinuous, and convex, then its Moreau-Yosida regularization E_τ is also convex. However, in the 2-Wasserstein metric, it is well-known that even when E satisfies analogous assumptions, E_τ is not always convex.[5] The key technical difference between the two metrics is that while

$$x \mapsto \frac{1}{2}||x - y||^2 \tag{1.12}$$

is 1-convex along geodesics,

$$\mu \mapsto \frac{1}{2} W_2^2(\mu, \nu) \tag{1.13}$$

is not λ-convex along geodesics, for any $\lambda \in \mathbb{R}$, if the dimension of the underlying space is greater than or equal to 2 [2, Example 9.1.5]. Since much of De Giorgi's "minimizing steps" approach to gradient flow relies on the 1-convexity of (1.12), this lack of convexity in the 2-Wasserstein case complicates the implementation of De Giorgi's scheme.

Ambrosio, Gigli, and Savaré circumvent this difficulty with their observation that, though $\mu \mapsto \frac{1}{2} W_2^2(\mu, \nu)$ is not 1-convex along all geodesics, it is 1-convex along a different class of curves. They define the set of *generalized geodesics* to be the union of these classes of curves over all $\nu \in \mathcal{P}(\mathbb{R}^d)$ (see Section 2.1). By considering functionals that are *convex along generalized geodesics*—a stronger condition than merely being convex along geodesics (see Section 2.2)—they deduce a priori estimates that provide detailed control over the gradient flow and discrete gradient flow.

[5] For the reader's convenience, we include an example in Section 3.

The key results that we will use concern functionals $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ that are proper, coercive, lower semicontinuous, and λ-convex along generalized geodesics (see Section 2.2).[6] With these assumptions, Ambrosio, Gigli, and Savaré show that if $\tau > 0$ is small enough so that $\lambda\tau > -1$, then for all $\mu \in \overline{D(E)}$ the proximal map

$$\mu \mapsto \mu_\tau \tag{1.14}$$

and the *discrete gradient flow* sequence

$$\mu_n = (\mu_{n-1})_\tau, \quad \mu_0 \in \overline{D(E)},$$

are well-defined. They go on to prove the 2-Wasserstein analogue of the exponential formula (1.2) relating the discrete gradient flow to the continuous gradient flow [2, Theorem 4.0.4]. If $\mu(t)$ is the solution to the continuous gradient flow of E with initial conditions $\mu(0) \in \overline{D(E)}$, then

$$\mu(t) = \lim_{n\to\infty} (J_{t/n})^n(\mu(0)) \ . \tag{1.15}$$

Using the assumption of convexity along generalized geodesics, Ambrosio, Gigli, and Savaré comprehensively develop the theory of continuous gradient flow. While this assumption is stronger than (standard) convexity along geodesics, it is not restrictive: all important examples of functionals that are convex along geodesics are also convex along generalized geodesics [2, Section 9.3].

In this paper, we take a closer look at the Moreau-Yosida regularization and the proximal map in the 2-Wasserstein metric for functionals that are convex along generalized geodesics. We show that, while the Moreau-Yosida regularization does not preserve E's convexity along all geodesics (as in the Hilbertian case), if E attains its minimum at $\bar{\mu}$, the Moreau-Yosida regularization does satisfy an "above the tangent line" inequality at $\bar{\mu}$. This type of inequality is a necessary condition for convexity—in particular, a function from $\mathbb{R}$ to $\mathbb{R}$ is convex if and only if it lies above its tangent line at every point.

1.1 THEOREM (Generalized convexity of E_τ)**.** *Given $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ proper, coercive, lower semicontinuous, and λ-convex along generalized geodesics with $\lambda \geq 0$, assume that E attains its minimum at $\bar{\mu}$. For $\tau > 0$, define $\lambda_\tau := \frac{\lambda}{1+\lambda\tau}$. Then for all $\mu \in \overline{D(E)}$, there exists a geodesic $\mu_\alpha^{\bar{\mu}\to\mu}$ from $\bar{\mu}$ to μ such that*

$$E_\tau(\mu_\alpha^{\bar{\mu}\to\mu}) \leq (1-\alpha)E_\tau(\bar{\mu}) + \alpha E_\tau(\mu) - \alpha(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\bar{\mu},\mu). \tag{1.16}$$

In Section 4.1, we show that (1.16) is sharp by presenting an example in which E is λ-convex and E_τ is no more than λ_τ-convex

As a consequence of Theorem 1.1, we show E_τ satisfies a Talagrand inequality and an HWI inequality.

1.2 THEOREM (Talagrand and HWI Inequalities)**.** *Under the assumptions of the Theorem 1.1, for all $\mu \in \overline{D(E)}$, we have the Talagrand inequality*

$$E_\tau(\mu) - E_\tau(\bar{\mu}) \geq \frac{\lambda_\tau}{2}W_2^2(\mu,\bar{\mu}) \tag{1.17}$$

[6] Note that Ambrosio, Gigli, and Savaré often state their results in the context when $\mu_0 = \delta_0$, the Dirac mass at the origin, so $\mathcal{P}_{\mu_0}(\mathbb{R}^d) = \mathcal{P}_2(\mathbb{R}^d)$. We quote their results in broader generality, since the proofs are easily adapted to this case.

and the HWI inequality

$$E_\tau(\mu) - E_\tau(\bar{\mu}) \leq |\nabla_W E_\tau(\mu)| W_2(\mu, \bar{\mu}) - \frac{\lambda_\tau}{2} W_2^2(\mu, \bar{\mu}) \ . \tag{1.18}$$

These inequalities capture E_τ's behavior at $\bar{\mu}$ from both ends of the "above the tangent line" inequality.

We also develop the analogy between Hilbertian metrics and the 2-Wasserstein metric by proving a contraction inequality for the proximal map. In a Hilbert space, if E is proper, lower semicontinuous, and convex, Moreau [16] showed that the proximal map satisfies

$$||x_\tau - y_\tau|| \leq ||x - y|| \ \forall x, y \in \mathcal{H}. \tag{1.19}$$

This turns out to be a rather miraculous property of the Hilbertian norm that fails even in simple Banach spaces. For example, consider the ℓ^∞ norm on $\mathbb{R}^2$. Fix two points $a = (0, 0)$ and $b = (1, 1)$, and let K be the closed half-space lying beneath the line $3x_2 = x_1 - 4$. Let E be the indicator function for K,

$$E(x) := \begin{cases} 0 & \text{if } x = (x_1, x_2) \in K \\ \infty & \text{otherwise.} \end{cases}$$

Then

$$J_\tau(y) := \operatorname*{argmin}_{x \in \mathbb{R}^2} \left\{ \frac{1}{2\tau} ||x - y||_\infty^2 + E(x) \right\} = \operatorname*{argmin}_{x \in K} \left\{ \frac{1}{2\tau} ||x - y||_\infty^2 \right\} \ .$$

Therefore, $J_\tau(a) = (1, -1)$ and $J_\tau(b) = (5/2, -1/2)$ for all $\tau > 0$. This is not a contraction since $||a - b||_\infty = 1 < 3/2 = ||J_\tau(a) - J_\tau(b)||_\infty$.

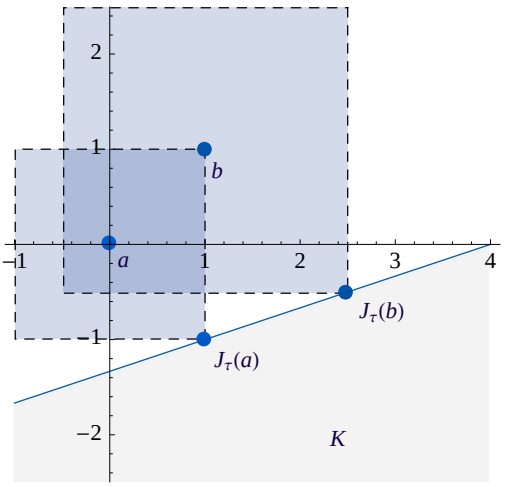

Figure 1: In the Banach space $\mathbb{R}^2$, endowed with the ℓ^∞ norm, the proximal map is not a contraction.

The situation for general metric spaces is even more involved than the situation for metrics induced by norms, and one does not expect a contraction to hold. Nevertheless, if E is appropriately convex, the continuous time gradient flow defined by (1.15) is contractive [2, Theorem 4.0.4] [19]. This gives hope that some contraction property of the proximal map is present at the discrete level and does not merely emerge in the limit.

Our next result shows that this is the case. In particular, we achieve contraction of the proximal map by making a small modification to the squared distance: given $\tau > 0$, we consider the functional $\Lambda_\tau : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ defined by

$$\Lambda_\tau(\mu, \nu) := W_2^2(\mu, \nu) + \frac{\tau^2}{2}|\nabla_W E(\mu)|^2 + \frac{\tau^2}{2}|\nabla_W E(\nu)|^2. \tag{1.20}$$

As before, we interpret $|\nabla_W E(\mu)|$ as the metric slope (1.11) when E and μ lack sufficient smoothness for the norm of the 2-Wasserstein gradient (1.9) to be well-defined.

Though we state the following theorem in the context of the 2-Wasserstein metric, it continues to hold in a more abstract setting: given a functional E on a complete metric space (X, d), if E is proper, coercive, lower semicontinuous, and satisfies [2, Assumption 4.0.1] for some $\lambda \in \mathbb{R}$, then the result remains true by replacing W_2 with d.

1.3 THEOREM (Contraction of proximal map)**.** *Given $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ proper, coercive, lower semicontinuous, and λ-convex along generalized geodesics, fix $\tau > 0$ small enough so that $\lambda\tau > -1$. Consider $\mu, \nu \in \overline{D(E)}$ and let $\Lambda_\tau : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ be given by (1.20). Then, if $\lambda \geq 0$, the proximal map is contracting in Λ_τ,*

$$\Lambda_\tau(\mu_\tau, \nu_\tau) \leq \Lambda_\tau(\mu, \nu). \tag{1.21}$$

More generally, for $\lambda \in \mathbb{R}$,

$$\begin{aligned}\Lambda_\tau(\mu_\tau, \nu_\tau) - \Lambda_\tau(\mu, \nu) \leq &-\frac{1}{2}(\tau|\nabla_W E(\nu)| - W_2(\nu, \nu_\tau))^2 - \frac{1}{2}(\tau|\nabla_W E(\mu)| - W_2(\mu, \mu_\tau))^2 \\ &- \frac{\lambda\tau}{2}\left[2W_2^2(\mu_\tau, \nu_\tau) + W_2^2(\mu, \nu_\tau) + W_2^2(\nu, \mu_\tau) + W_2^2(\nu, \nu_\tau) + W_2^2(\mu, \mu_\tau)\right] .\end{aligned} \tag{1.22}$$

In Section 4.1, we show that the inequality (1.22) is sharp. Then, in Section 4.2, we apply (1.21) together with scaling properties of the W_2 metric to derive sharp polynomial rates of convergence to Barenblatt profiles for certain fast diffusion and porous medium equations. Otto originally deduced these results in [19] by considering a modified gradient flow problem for λ-convex functionals with $\lambda > 0$. The contraction inequality (1.21) provides a simple route to such results. The fast diffusion and porous media equations also provide examples of strictly convex functionals for which the proximal map is strictly contracting in Λ_τ but not in W_2.

1.4 Remark. While Ambrosio, Gigli, and Savaré do not explicitly consider monotonicity results for modifications of the squared distance along the discrete gradient flow, such a result (for a different modification) can be found by reading between the lines in [2, Lemma 4.2.4]. Consider the alternative modification to the squared distance function defined by

$$\widetilde{\Lambda}_\tau(\mu, \nu) := W_2^2(\mu, \nu) + \tau E(\mu) + \tau E(\nu) . \tag{1.23}$$

If one takes the final inequality on [2, page 92] for $\lambda = 0$ and $n = 1$, rearranges terms, and symmetrizes in μ and ν, one obtains (1.21) with $\widetilde{\Lambda}_\tau$ in place of Λ_τ. A key difference between $\widetilde{\Lambda}_\tau$ and our functional Λ_τ is that, for measures μ and ν with $|\nabla_W E(\mu)|$ and $|\nabla_W E(\mu)| < \infty$, Λ_τ involves only an $\mathcal{O}(\tau^2)$ correction to $W_2^2(\mu, \nu)$, while $\widetilde{\Lambda}_\tau$ involves an $\mathcal{O}(\tau)$ correction to $W_2^2(\mu, \nu)$.

1.5 Remark. While one might first suppose that Λ_τ could only be used to study discrete gradient flows with initial data μ, ν satisfying $|\nabla_W E(\mu)|, |\nabla_W E(\mu)| < \infty$, when E is strictly convex, the discrete gradient flow produces this regularity in one step (see Lemma 2.2). We shall see an example of this in Section 4.2 when we apply Theorem 1.3 to the discrete gradient flow for the Rényi entropies.

For $\lambda > 0$, one can extract from (1.22) a useful inequality that implies, among other things, an optimal exponential rate of decrease of $\Lambda_\tau(\mu, \bar{\mu})$ when E has a minimizer $\bar{\mu}$ (necessarily unique due to the strict convexity).

1.6 COROLLARY (The case $\lambda > 0$)**.** *Consider $\lambda > 0$ and $\tau > 0$ sufficiently small so that $\tau\lambda \leq 1$. Then for all E satisfying the hypotheses of Theorem 1.3 and $\mu, \nu \in \overline{D(E)}$,*

$$(1 + \tau\lambda)\Lambda_\tau(\mu_\tau, \nu_\tau) \leq (1 - \tau\lambda)\Lambda_\tau(\mu, \nu) + 3\lambda\tau\Lambda_\tau^{1/2}(\mu, \nu)[W_2(\mu, \mu_\tau) + W_2(\nu, \nu_\tau)] \ . \tag{1.24}$$

We give the proof of this corollary in Section 3. However, to explain its consequences, we state and prove a simple discrete Gronwall type inequality. It is a discrete version of the continuous time inequality [2, Lemma 4.1.8]. (See [3] and [9] for related discrete Gronwall inequalities.)

1.7 LEMMA (A discrete Gronwall type inequality)**.** *Let $\lambda, \tau > 0$, and let $\{a_n\}$ and $\{b_n\}$ be two sequences of non-negative numbers such that for all $n \geq 0$,*

$$(1 + \tau\lambda)a_n \leq (1 - \tau\lambda)a_{n-1} + \tau a_{n-1}^{1/2} b_n \ . \tag{1.25}$$

Then,

$$a_n^{1/2} \leq (1 + \lambda\tau)^{-n} a_0^{1/2} + \sqrt{\frac{\tau}{2\lambda}}(1 + \lambda\tau) \left(\sum_{k=1}^{n} b_k^2 \right)^{1/2} \ .$$

Consider the discrete gradient flow of E starting from $\mu \in D(E)$ with $\tau > 0$ and $\tau\lambda \leq 1$. Let $\mu_0 := \mu$ and inductively define $\{\mu_n\}$ by repeated application of the proximal map. Define $\{\nu_n\}$ in the same way, starting from $\nu \in D(E)$. Now, apply Lemma 1.7 and Corollary 1.6 to these discrete gradient flows of E, taking

$$a_n := \Lambda_\tau(\mu_n, \nu_n) \qquad \text{and} \qquad b_n := 3\lambda\sqrt{2W_2^2(\mu_{n-1}, \mu_n) + 2W_2^2(\nu_{n-1}, \nu_n)} \ .$$

Since

$$W_2^2(\mu, \mu_\tau) \leq 2\tau[E(\mu) - E(\mu_\tau)] \ , \tag{1.26}$$

$\sum_{k=1}^n b_k^2$ is bounded by a telescoping sum: $\sum_{k=1}^n b_k^2 \leq \tau 36\lambda^2[(E(\mu) - E(\mu_n)) + (E(\nu) - E(\nu_n))]$. In case E is bounded below, we may assume without loss of generality that E is non-negative. Then,

$$\Lambda_\tau^{1/2}(\mu_n, \nu_n) \leq (1 + \lambda\tau)^{-n}\Lambda_\tau^{1/2}(\mu, \nu) + \lambda\tau\frac{6(1 + \lambda\tau)}{\sqrt{2\lambda}}\sqrt{E(\mu) + E(\nu)} \ . \tag{1.27}$$

Thus, for positive λ and sufficiently small τ, $\Lambda_\tau^{1/2}(\mu_n, \nu_n)$ decays "exponentially fast" at rate λ up to the time that this quantity becomes $\mathcal{O}(\tau)$.[7]

The proof of Lemma 1.7 is elementary, so we provide it here, closing this section.

[7] At this point, we may use the bound $E(\mu_n) \leq (1 + \lambda\tau)^{-2n} E(\mu)$ [2, Theorem 3.1.6] and apply (1.27) iteratively.

Proof of Lemma 1.7. Multiply both sides of (1.25) by $(1+\tau\lambda)^{2n-1}$ to obtain

$$(1+\tau\lambda)^{2n}a_n \le (1-(\tau\lambda)^2)(1-\tau\lambda)^{2n-2}a_{n-1} + \tau\left((1+\tau\lambda)^{2n-2}a_{n-1}\right)^{1/2}(1+\tau\lambda)^n b_n \ .$$

Defining

$$\widetilde{a}_n := (1+\tau\lambda)^{2n}a_n \qquad \text{and} \qquad \widetilde{b}_n := \tau(1+\tau\lambda)^n b_n \ ,$$

we have $\widetilde{a}_n \le \widetilde{a}_{n-1} + \widetilde{a}_{n-1}^{1/2}\widetilde{b}_n$, and therefore $\widetilde{a}_n \le a_0 + \sum_{k=1}^{n}\widetilde{a}_{k-1}^{1/2}\widetilde{b}_k$. Defining

$$c_n := \max\{\widetilde{a}_k \ : \ 0 \le k \le n\} \ ,$$

we have $c_n \le a_0 + c_n^{1/2}\sum_{k=1}^{n}\widetilde{b}_k$. This quadratic inequality implies that $c_n^{1/2} \le a_0^{1/2} + \sum_{k=1}^{n}\widetilde{b}_k$. By the Cauchy-Schwarz inequality, and the fact that for $\alpha := (1+\lambda\tau)^2 \ge 1$, $\sum_{k=1}^{n}\alpha^k \le \frac{\alpha}{\alpha-1}\alpha^n$,

$$\sum_{k=1}^{n}\widetilde{b}_k \le \frac{\sqrt{\tau}(1+\lambda\tau)^{n+1}}{\sqrt{2\lambda}}\left(\sum_{k=1}^{n}b_k^2\right)^{1/2} \ .$$

□

2 Generalized Convexity and the Proximal Map

2.1 Generalized Geodesics

In a Hilbert space, $x \mapsto \frac{1}{2}||x-y||^2$ is 1-convex along geodesics. However, the same is not true for the squared 2-Wasserstein distance when the dimension of the underlying space is greater than or equal to 2 [2, Example 9.1.5]. Instead, Ambrosio, Gigli, and Savaré observe that $\mu \mapsto \frac{1}{2}W_2^2(\mu,\nu)$ is convex along a different set of curves, which we now describe.

Fix $\mu_1,\mu_2,\mu_3 \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)$ with optimal plans $\boldsymbol{\mu}_{1,2} \in \Gamma_0(\mu_1,\mu_2)$, $\boldsymbol{\mu}_{1,3} \in \Gamma_0(\mu_1,\mu_3)$. For $1 \le i < j \le 3$, let $\pi_{i,j}$ be the projection onto the ith and jth components of $\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d$. Fix $\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d)$ so that $\pi_{1,2}\#\boldsymbol{\mu} = \boldsymbol{\mu}_{1,2}$ and $\pi_{1,3}\#\boldsymbol{\mu} = \boldsymbol{\mu}_{1,3}$ [2, Lemma 5.3.2]. (We use bold font to distinguish probability measures on $\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d$ or $\mathbb{R}^d\times\mathbb{R}^d$ from probability measures on $\mathbb{R}^d$.) As in [2, Definition 9.2.2], a *generalized geodesic joining μ_2 to μ_3 with base μ_1* is a curve of the form

$$\mu_\alpha^{2\to3} : [0,1] \to \mathcal{P}(\mathbb{R}^d), \quad \mu_\alpha^{2\to3} := \left((1-\alpha)\pi_2 + \alpha\pi_3\right)\#\boldsymbol{\mu}.$$

In the case $\mu_1 \in P_2^a(\mathbb{R}^d)$ and $\mu_2,\mu_3 \in \mathcal{P}_2(\mathbb{R}^d)$, this reduces to

$$\mu_\alpha^{2\to3} : [0,1] \to \mathcal{P}(\mathbb{R}^d), \quad \mu_\alpha^{2\to3} = \left((1-\alpha)\mathbf{t}_{\mu_1}^{\mu_2} + \alpha\mathbf{t}_{\mu_1}^{\mu_3}\right)\#\mu_1.$$

Ambrosio, Gigli, and Savaré demonstrate that $\mu \mapsto \frac{1}{2}W_2^2(\mu,\mu_1)$ is 1-convex along any generalized geodesic $\mu_\alpha^{2\to3}$ with base μ_1, for all $\mu^2,\mu^3 \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)$ [2, Lemma 9.2.1]. Note that if the base μ_1 equals either μ_2 or μ_3, $\mu_\alpha^{2\to3}$ is a (standard) geodesic joining μ_2 and μ_3. Thus, while $\mu \to \frac{1}{2}W_2^2(\mu,\mu_1)$ is not convex along geodesics (in the sense that it is not convex along *all* geodesics), it is convex along *some* geodesics.

2.2 Functionals $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$

Fix a Borel probability measure μ_0. We consider functionals $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ that satisfy the following conditions:

- *proper:* $D(E) := \{\mu \in \mathcal{P}_{\mu_0}(\mathbb{R}^d) : E(\mu) < \infty\} \neq \emptyset$
- *coercive*[8]: There exists $\tau^* > 0$ such that for all $0 < \tau < \tau^*, \mu \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)$,

$$E_\tau(\mu) = \inf_{\nu \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)} \left\{ \frac{1}{2\tau} W_2^2(\mu, \nu) + E(\nu) \right\} > -\infty.$$

  As noted in [2, Lemma 2.2.1], by a triangle inequality argument, it is enough to check that there exists $\tau_0 > 0$ such that

$$E_{\tau_0}(\mu_0) = \inf_{\nu \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)} \left\{ \frac{1}{2\tau_0} W_2^2(\mu_0, \nu) + E(\nu) \right\} > -\infty. \tag{2.1}$$

- *lower semicontinuous:* For all $\mu_n, \mu \in P_{\mu_0}(\mathbb{R}^d)$ such that $\mu_n \to \mu$ in W_2,

$$\liminf_{n\to\infty} E(\mu_n) \geq E(\mu).$$

- λ-*convex along generalized geodesics*: For any $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)$, there exists a generalized geodesic $\mu_\alpha^{2\to 3}$ from μ_2 to μ_3 with base μ_1 such that for all $\alpha \in [0, 1]$,

$$E(\mu_\alpha^{2\to 3}) \leq (1-\alpha)E(\mu_2) + \alpha E(\mu_3) - \alpha(1-\alpha)\frac{\lambda}{2} \int |x_2 - x_3|^2 d\boldsymbol{\mu}(x). \tag{2.2}$$

  Note that, for $\lambda > 0$, this condition is stronger than requiring that $E(\mu_\alpha^{2\to 3})$, considered as a real-valued function of $\alpha \in [0, 1]$, be $\lambda W_2^2(\mu_2, \mu_3)$ convex, since

$$\int |x_2 - x_3|^2 d\boldsymbol{\mu} \geq W_2^2(\mu_2, \mu_3).$$

If E is λ-convex along generalized geodesics, then in particular it is λ-*convex:* for any $\mu_1, \mu_2 \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)$, there exists a geodesic $\mu_\alpha^{1\to 2}$ from μ_1 to μ_2 such that for all $\alpha \in [0, 1]$,

$$E(\mu_\alpha^{1\to 2}) \leq (1-\alpha)E(\mu_1) + \alpha E(\mu_2) - \alpha(1-\alpha)\frac{\lambda}{2} W_2^2(\mu_1, \mu_2).$$

This is equivalent to $E(\mu_\alpha^{1\to 2})$, considered as a real-valued function of $\alpha \in [0, 1]$, being $\lambda W_2^2(\mu_1, \mu_2)$ convex [2, Remark 9.1.2].

The requirement that a functional $E : \mathcal{P}_{\mu_0} \to \mathbb{R} \cup \{\infty\}$ be proper, coercive, lower semicontinuous, and convex along generalized geodesics is the natural analogue of the Hilbertian requirement that

[8] In the case $\mu_0 = \delta_0$, the Dirac mass at the origin, this is equivalent to the definition of coercivity in [2], where Ambrosio, Gigli, and Savaré require that there exist some $\tau_* > 0$ and $\mu_* \in \mathcal{P}_2(\mathbb{R}^d)$ such that

$$\inf_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \frac{1}{2\tau_*} W_2^2(\mu_*, \nu) + E(\nu) \right\} > -\infty.$$

a functional $E : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ be proper, lower semicontinuous, and convex. The two differences are the addition of the coercivity assumption and the strengthening of the convexity assumption. In a Hilbert space $\mathcal{H}$, all functionals that are proper, lower semicontinuous, and convex are also coercive (in this sense), so the addition of the coercivity assumption is a natural way to ensure that the 2-Wasserstein Moreau-Yosida regularization is not identically $-\infty$. The convexity assumption is strengthened because convexity along generalized geodesics is the useful 2-Wasserstein analogue of Hilbertian convexity. While in a Hilbert space, $x \mapsto \frac{1}{2}||x-y||^2$ is 1-convex along all geodesics, the same does not hold for the 2-Wasserstein metric. Requiring convexity of the functional on a larger class of curves compensates for the weaker convexity W_2^2.

2.3 Further Results About the Proximal Map

In the following theorem, we collect some key results from [2, Theorem 4.1.2, Corollary 4.1.3] regarding the proximal map.

2.1 THEOREM. *Given $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ proper, coercive, lower semicontinuous, and λ-convex along generalized geodesics, fix $\tau > 0$ small enough so that $\tau\lambda > -1$. Then, for $\mu \in \overline{D(E)}$, the proximal map*

$$\mu \mapsto \mu_\tau$$

is well-defined. Furthermore, the following variational inequality holds:

$$\frac{1}{2\tau}\left(W_2^2(\mu_\tau,\nu) - W_2^2(\mu,\nu)\right) + \frac{\lambda}{2}W_2^2(\mu_\tau,\nu) \le E(\nu) - E(\mu_\tau) - \frac{1}{2\tau}W_2^2(\mu,\mu_\tau), \quad \forall \nu \in D(E). \quad (2.3)$$

When the proximal map is well-defined, it satisfies an Euler-Lagrange equation—a fact originally observed by Otto in [18, 19]. We state this result in the framework of [2, Lemma 10.1.2].

2.2 LEMMA. *Given $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ proper, coercive, lower semicontinuous, and λ-convex along generalized geodesics, fix $\tau > 0$ small enough so that $\tau\lambda > -1$. Assume that $\mu \in \overline{D(E)}$ so $\mu \mapsto \mu_\tau$ is well-defined by Theorem 2.1. Then,*

$$\tau|\nabla_W E(\mu_\tau)| \le W_2(\mu,\mu_\tau). \quad (2.4)$$

We may interpret $|\nabla_W E(\mu_\tau)|$ as the metric slope (1.11) when E and μ lack sufficient smoothness for the norm of the 2-Wasserstein gradient (1.9) to be well-defined.

On the other hand, if $\mu \in \mathcal{P}_2^a(\mathbb{R}^d)$ and both E and μ_τ are smooth enough so that the 2-Wasserstein gradient $\nabla_W E(\mu_\tau)$ is well-defined by (1.8), then

$$\mathbf{t}_{\mu_\tau}^{\mu} = \mathrm{id} + \tau\nabla\frac{\delta E}{\delta\rho}(\mu_\tau) \quad (2.5)$$

μ_τ-almost everywhere and

$$\tau|\nabla_W E(\mu_\tau)| = W_2(\mu,\mu_\tau). \quad (2.6)$$

Proof. (2.4) follows from [2, Theorem 3.1.6].

(2.5) follows from [2, Lemma 10.1.2] and the fact that, when E is differentiable, $\nabla\frac{\delta E}{\delta\rho}(\mu_\tau)$ is the unique element of its subdifferential at μ_τ.

(2.6) follows from (2.5) by considering the $L^2(\mu_\tau)$ norm of $\mathbf{t}_{\mu_\tau}^{\mu} - \mathrm{id} = \tau\nabla\frac{\delta E}{\delta\rho}(\mu_\tau)$.

□

3 Proofs of Theorems 1.1, 1.2, and 1.3 and Corollary 1.6

We now prove the theorems and corollaries announced in the introduction, turning first to the generalized convexity of E_τ. In a Hilbert space, if E is proper, lower semicontinuous, and convex, then its Moreau-Yosida regularization E_τ is also convex. It is well known that the exact analogue in the 2-Wasserstein metric is false. For lack of a reference, we provide the following example.

Fix $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and define $E : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ by

$$E(\mu) := \begin{cases} 0 & \text{if } \mu = \mu_0 \\ \infty & \text{otherwise.} \end{cases} \tag{3.1}$$

E is proper, coercive, lower semicontinuous, and convex along all curves in $\mathcal{P}_2(\mathbb{R}^d)$. In particular, E is convex along generalized geodesics. By definition,

$$E_\tau(\mu) = \inf_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \frac{1}{2\tau} W_2^2(\mu, \nu) + E(\nu) \right\} = \frac{1}{2\tau} W_2^2(\mu, \mu_0) \ .$$

By [2, Example 9.1.5], when the dimension of the underlying space satisfies $d \geq 2$, E_τ is not λ-convex along geodesics for any $\lambda \in \mathbb{R}$.

As demonstrated by the previous example, the convexity of E_τ is related to the convexity of the squared 2-Wasserstein distance. This also holds in the Hilbertian case, where the convexity of E_τ is a consequence of the 1-convexity of $x \mapsto \frac{1}{2}||x - y||^2$ [17]. Therefore, it is natural that our proof of the convexity inequality for E_τ requires the following convexity inequality for W_2^2.

3.1 LEMMA (Convexity inequality for W_2^2)**.** *Fix three measures $\mu_1, \mu_2, \mu_3 \in \mathcal{P}(\mathbb{R}^d)$ that are a finite 2-Wasserstein distance apart. Let $\mu_\alpha^{1\to 3}$ be a generalized geodesic from μ_1 to μ_3 with base point μ_2,*

$$\mu_\alpha^{1\to 3} := ((1-\alpha)\pi_1 + \alpha\pi_3)\#\boldsymbol{\mu} \ ,$$

where $\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ satisfies $\boldsymbol{\mu}_{1,2} := \pi_{1,2}\#\boldsymbol{\mu} \in \Gamma_0(\mu_1, \mu_2)$ and $\boldsymbol{\mu}_{2,3} := \pi_{2,3}\#\boldsymbol{\mu} \in \Gamma_0(\mu_2, \mu_3)$. Let $\mu_\alpha^{1\to 2}$ be the geodesic from μ_1 to μ_2 defined by

$$\mu_\alpha^{1\to 2} := ((1-\alpha)\pi_1 + \alpha\pi_2)\,\#\boldsymbol{\mu}_{1,2} \ .$$

Then,

$$W_2^2(\mu_\alpha^{1\to 2}, \mu_\alpha^{1\to 3}) \leq (1-\alpha)W_2^2(\mu_1, \mu_1) + \alpha W_2^2(\mu_2, \mu_3) - \alpha(1-\alpha)W_2^2(\mu_2, \mu_3). \tag{3.2}$$

Proof. Note that

$$\mu_\alpha^{1\to 2} = ((1-\alpha)\pi_1 + \alpha\pi_2)\,\#\boldsymbol{\mu}_{1,2} = ((1-\alpha)\pi_1 + \alpha\pi_2)\,\#\boldsymbol{\mu}.$$

Then by [2, Equation 7.1.6],

$$\begin{aligned} W_2^2(\mu_\alpha^{1\to 2}, \mu_\alpha^{1\to 3}) &\leq \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} |[(1-\alpha)\pi_1 + \alpha\pi_3] - [(1-\alpha)\pi_1 + \alpha\pi_2]|^2 \, d\boldsymbol{\mu} \\ &= \alpha^2 \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} |\pi_2 - \pi_3|^2 \, d\boldsymbol{\mu} \\ &= \alpha^2 \int_{\mathbb{R}^d\times\mathbb{R}^d} |\pi_2 - \pi_3|^2 \, d\boldsymbol{\mu}_{2,3} \\ &= \alpha^2 W_2^2(\mu_2, \mu_3) \\ &= (1-\alpha)W_2^2(\mu_1, \mu_1) + \alpha W_2^2(\mu_2, \mu_3) - \alpha(1-\alpha)W_2^2(\mu_2, \mu_3). \end{aligned}$$

□

We now use this convexity inequality for W_2^2 to prove Theorem 1.1.

Proof of Theorem 1.1. Since E is proper, coercive, lower semicontinuous, and λ-convex along generalized geodesics for $\lambda \geq 0$, by Theorem 2.1, the proximal map $\mu \mapsto \mu_\tau$ is well-defined for $\mu \in \overline{D(E)}$ and $\tau > 0$. Let $\mu_\alpha^{\bar{\mu}\to\mu_\tau}$ be the generalized geodesic from $\bar{\mu}$ to μ_τ with base point μ on which E satisfies equation (2.2). Defining $\mu_1 := \bar{\mu}$, $\mu_2 := \mu$, and $\mu_3 := \mu_\tau$, let $\mu_\alpha^{\bar{\mu}\to\mu}$ be the geodesic from $\bar{\mu}$ to μ described in Lemma 3.1. By Lemma 3.1,

$$W_2^2(\mu_\alpha^{\bar{\mu}\to\mu}, \mu_\alpha^{\bar{\mu}\to\mu_\tau}) \leq (1-\alpha)W_2^2(\bar{\mu},\bar{\mu}) + \alpha W_2^2(\mu,\mu_\tau) - \alpha(1-\alpha)W_2^2(\mu,\mu_\tau).$$

This allows us to bound $E_\tau(\mu_\alpha^{\bar{\mu}\to\mu})$ from above:

$$\begin{aligned}
E_\tau(\mu_\alpha^{\bar{\mu}\to\mu}) &= \inf_{\nu\in\mathcal{P}_{\mu_0}(\mathbb{R}^d)} \left\{ \frac{1}{2\tau} W_2^2(\mu_\alpha^{\bar{\mu}\to\mu}, \nu) + E(\nu) \right\} \\
&\leq \frac{1}{2\tau} W_2^2(\mu_\alpha^{\bar{\mu}\to\mu}, \mu_\alpha^{\bar{\mu}\to\mu_\tau}) + E(\mu_\alpha^{\bar{\mu}\to\mu_\tau}) \\
&\leq \frac{1}{2\tau}\left((1-\alpha)W_2^2(\bar{\mu},\bar{\mu}) + \alpha W_2^2(\mu,\mu_\tau) - \alpha(1-\alpha)W_2^2(\mu,\mu_\tau)\right) \\
&\quad + (1-\alpha)E(\bar{\mu}) + \alpha E(\mu_\tau) - \alpha(1-\alpha)\frac{\lambda}{2}W_2^2(\bar{\mu},\mu_\tau) \\
&\leq (1-\alpha)E_\tau(\bar{\mu}) + \alpha E_\tau(\mu) - \alpha(1-\alpha)\left(\frac{1}{2\tau}W_2^2(\mu,\mu_\tau) + \frac{\lambda}{2}W_2^2(\bar{\mu},\mu_\tau)\right) .
\end{aligned}$$

In the last step, we used that $(\bar{\mu})_\tau = \bar{\mu}$, since E attains its minimum at $\bar{\mu}$. Now, we apply

$$\alpha a^2 + \beta b^2 \geq \frac{\alpha\beta}{\alpha+\beta}(a+b)^2 \ , \text{ for } \alpha > 0, \beta \geq 0$$

with $\alpha = 1/\tau$ and $\beta = \lambda$:

$$\begin{aligned}
E_\tau(\mu_\alpha^{\bar{\mu}\to\mu}) &\leq (1-\alpha)\left(\frac{1}{2\tau}W_2^2(\bar{\mu},\bar{\mu}) + E(\bar{\mu})\right) + \alpha\left(\frac{1}{2\tau}W_2^2(\mu,\mu_\tau) + E(\mu_\tau)\right) \\
&\quad - \alpha(1-\alpha)\frac{\lambda_\tau}{2}\left(W_2(\mu,\mu_\tau) + W_2(\bar{\mu},\mu_\tau)\right)^2 \\
&\leq (1-\alpha)\left(\frac{1}{2\tau}W_2^2(\bar{\mu},\bar{\mu}) + E(\bar{\mu})\right) + \alpha\left(\frac{1}{2\tau}W_2^2(\mu,\mu_\tau) + E(\mu_\tau)\right) - \alpha(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\mu,\bar{\mu}).
\end{aligned}$$

Finally, since E attains its minimum at $\bar{\mu}$, $(\bar{\mu})_\tau = \bar{\mu}$. Therefore,

$$\begin{aligned}
E_\tau(\mu_\alpha^{\bar{\mu}\to\mu}) &\leq (1-\alpha)\left(\frac{1}{2\tau}W_2^2(\bar{\mu},(\bar{\mu})_\tau) + E((\bar{\mu})_\tau)\right) + \alpha\left(\frac{1}{2\tau}W_2^2(\mu,\mu_\tau) + E(\mu_\tau)\right) - \alpha(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\mu,\bar{\mu}) \\
&= (1-\alpha)E_\tau(\bar{\mu}) + \alpha E_\tau(\mu) - \alpha(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\mu,\bar{\mu}).
\end{aligned}$$

□

We now use this convexity inequality to prove Theorem 1.2.

Proof of Theorem 1.2. We first prove the Talagrand inequality. Since E attains its minimum at $\bar{\mu}$, so does E_τ. Therefore, (1.16) implies that, for all $\mu \in \overline{D(E)}$,

$$E_\tau(\bar{\mu}) \le E_\tau(\mu_\alpha^{\bar{\mu}\to\mu}) \le (1-\alpha)E_\tau(\bar{\mu}) + \alpha E_\tau(\mu) - \alpha(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\bar{\mu},\mu) \ .$$

Rearranging gives

$$\alpha(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\bar{\mu},\mu) \le \alpha\left(E_\tau(\mu) - E_\tau(\bar{\mu})\right) \ .$$

Thus, for all $\alpha \in (0,1)$,

$$(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\bar{\mu},\mu) \le E_\tau(\mu) - E_\tau(\bar{\mu}) \ .$$

Sending $\alpha \to 0$ gives the Talagrand inequality (1.17).

We now prove the HWI inequality. Again by (1.16), for all $\mu \in \overline{D(E)}$,

$$E_\tau(\mu_\alpha^{\bar{\mu}\to\mu}) \le (1-\alpha)E_\tau(\bar{\mu}) + \alpha E_\tau(\mu) - \alpha(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\mu,\bar{\mu}) \ .$$

Rearranging and using $\mu_\alpha^{\bar{\mu}\to\mu} = \mu_{1-\alpha}^{\mu\to\bar{\mu}}$ and $(1-\alpha)W_2(\mu,\bar{\mu}) = W_2(\mu,\mu_{1-\alpha}^{\mu\to\bar{\mu}})$ gives, for $\alpha \in (0,1)$,

$$\begin{aligned}
(1-\alpha)E_\tau(\mu) - (1-\alpha)E_\tau(\bar{\mu}) &\le E_\tau(\mu) - E_\tau(\mu_{1-\alpha}^{\mu\to\bar{\mu}}) - \alpha(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\mu,\bar{\mu}) \\
E_\tau(\mu) - E_\tau(\bar{\mu}) &\le \frac{E_\tau(\mu) - E_\tau(\mu_{1-\alpha}^{\mu\to\bar{\mu}})}{1-\alpha} - \alpha\frac{\lambda_\tau}{2}W_2^2(\mu,\bar{\mu}) \\
E_\tau(\mu) - E_\tau(\bar{\mu}) &\le \frac{E_\tau(\mu) - E_\tau(\mu_{1-\alpha}^{\mu\to\bar{\mu}})}{W_2(\mu,\mu_{1-\alpha}^{\mu\to\bar{\mu}})}W_2(\mu,\bar{\mu}) - \alpha\frac{\lambda_\tau}{2}W_2^2(\mu,\bar{\mu}) \ .
\end{aligned}$$

Sending $\alpha \to 1$ gives the HWI Inequality (1.18). □

Finally, we turn to the proof of Theorem 1.3.

Proof of Theorem 1.3. By Theorem 2.1, replacing ν with ν_τ,

$$\frac{1}{2\tau}\left(W_2^2(\mu_\tau,\nu_\tau) - W_2^2(\mu,\nu_\tau)\right) + \frac{\lambda}{2}W_2^2(\mu_\tau,\nu_\tau) \le E(\nu_\tau) - E(\mu_\tau) - \frac{1}{2\tau}W_2^2(\mu,\mu_\tau).$$

Similarly,

$$\frac{1}{2\tau}\left(W_2^2(\nu_\tau,\mu) - W_2^2(\nu,\mu)\right) + \frac{\lambda}{2}W_2^2(\nu_\tau,\mu) \le E(\mu) - E(\nu_\tau) - \frac{1}{2\tau}W_2^2(\nu,\nu_\tau).$$

Adding these and multiplying by 2τ gives

$$W_2^2(\mu_\tau,\nu_\tau) - W_2^2(\nu,\mu) + \lambda\tau\left[W_2^2(\mu_\tau,\nu_\tau) + W_2^2(\mu,\nu_\tau)\right] \le 2\tau\left[E(\mu) - E(\mu_\tau)\right] - W_2^2(\mu,\mu_\tau) - W_2^2(\nu,\nu_\tau).$$

Symmetrically, we also have

$$W_2^2(\mu_\tau,\nu_\tau) - W_2^2(\nu,\mu) + \lambda\tau\left[W_2^2(\mu_\tau,\nu_\tau) + W_2^2(\nu,\mu_\tau)\right] \quad \le 2\tau\left[E(\nu) - E(\nu_\tau)\right] - W_2^2(\mu,\mu_\tau) - W_2^2(\nu,\nu_\tau).$$

Averaging gives

$$W_2^2(\mu_\tau,\nu_\tau) - W_2^2(\nu,\mu) + \frac{\lambda\tau}{2}\left[2W_2^2(\mu_\tau,\nu_\tau) + W_2^2(\mu,\nu_\tau) + W_2^2(\nu,\mu_\tau)\right]$$
$$\leq \tau\left[E(\nu) - E(\nu_\tau) + E(\mu) - E(\mu_\tau)\right] - W_2^2(\mu,\mu_\tau) - W_2^2(\nu,\nu_\tau).$$

This allows us to bound the change in $\Lambda_\tau(\mu,\nu)$ from above:

$$\begin{aligned}
\Lambda_\tau(\mu_\tau,\nu_\tau) - \Lambda_\tau(\mu,\nu) &= W_2^2(\mu_\tau,\nu_\tau) + \frac{\tau^2}{2}|\nabla_W E(\mu_\tau)|^2 + \frac{\tau^2}{2}|\nabla_W E(\nu_\tau)|^2 \\
&\quad -W_2^2(\mu,\nu) - \frac{\tau^2}{2}|\nabla_W E(\mu)|^2 - \frac{\tau^2}{2}|\nabla_W E(\nu)|^2 \\
&\leq \tau\left[E(\nu) - E(\nu_\tau) + E(\mu) - E(\mu_\tau)\right] - W_2^2(\mu,\mu_\tau) - W_2^2(\nu,\nu_\tau) \\
&\quad +\frac{\tau^2}{2}|\nabla_W E(\mu_\tau)|^2 + \frac{\tau^2}{2}|\nabla_W E(\nu_\tau)|^2 - \frac{\tau^2}{2}|\nabla_W E(\mu)|^2 - \frac{\tau^2}{2}|\nabla_W E(\nu)|^2 \\
&\quad -\frac{\lambda\tau}{2}\left[2W_2^2(\mu_\tau,\nu_\tau) + W_2^2(\mu,\nu_\tau) + W_2^2(\nu,\mu_\tau)\right].
\end{aligned}$$

By [2, Equation 10.1.7, Lemma 10.1.5] and Hölder's inequality, the λ-convexity of E implies

$$E(\nu) - E(\nu_\tau) \leq |\nabla_W E(\nu)| W_2(\nu,\nu_\tau) - \frac{\lambda}{2}W_2^2(\nu,\nu_\tau)\ . \tag{3.3}$$

Combining this with the Euler-Lagrange equation (2.4),

$$\begin{aligned}
\Lambda_\tau(\mu_\tau,\nu_\tau) - \Lambda_\tau(\mu,\nu) &\leq \tau|\nabla_W E(\nu)| W_2(\nu,\nu_\tau) + \tau|\nabla_W E(\mu)| W_2(\mu,\mu_\tau) - W_2^2(\mu,\mu_\tau) - W_2^2(\nu,\nu_\tau) \\
&\quad +\frac{1}{2}W_2^2(\mu,\mu_\tau) + \frac{1}{2}W_2^2(\nu,\nu_\tau) - \frac{\tau^2}{2}|\nabla_W E(\mu)|^2 - \frac{\tau^2}{2}|\nabla_W E(\nu)|^2 \\
&\quad -\frac{\lambda\tau}{2}\left[2W_2^2(\mu_\tau,\nu_\tau) + W_2^2(\mu,\nu_\tau) + W_2^2(\nu,\mu_\tau)\right] - \frac{\lambda\tau}{2}\left[W_2^2(\nu,\nu_\tau) + W_2^2(\mu,\mu_\tau)\right].
\end{aligned}$$

Completing the square gives the result:

$$\begin{aligned}
\Lambda_\tau(\mu_\tau,\nu_\tau) - \Lambda_\tau(\mu,\nu) &\leq -\frac{1}{2}(\tau|\nabla_W E(\nu)| - W_2(\nu,\nu_\tau))^2 - \frac{1}{2}(\tau|\nabla_W E(\mu)| - W_2(\mu,\mu_\tau))^2 \\
&\quad -\frac{\lambda\tau}{2}\left[2W_2^2(\mu_\tau,\nu_\tau) + W_2^2(\mu,\nu_\tau) + W_2^2(\nu,\mu_\tau) + W_2^2(\nu,\nu_\tau) + W_2^2(\mu,\mu_\tau)\right]
\end{aligned}$$

□

Proof of Corollary 1.6. First, we use $\lambda > 0$ and the Euler-Lagrange equation (2.4) to rewrite (1.22):

$$\begin{aligned}
\Lambda_\tau(\mu_\tau,\nu_\tau) - \Lambda_\tau(\mu,\nu) &\leq -\frac{1}{2}(\tau|\nabla_W E(\nu)| - W_2(\nu,\nu_\tau))^2 - \frac{1}{2}(\tau|\nabla_W E(\mu)| - W_2(\mu,\mu_\tau))^2 \\
&\quad -\frac{\lambda\tau}{2}\left[2W_2^2(\mu_\tau,\nu_\tau) + W_2^2(\mu,\nu_\tau) + W_2^2(\nu,\mu_\tau) + \tau^2|\nabla_W E(\nu_\tau)|^2 + \tau^2|\nabla_W E(\mu_\tau)|^2\right] \\
&= -\frac{1}{2}(\tau|\nabla_W E(\nu)| - W_2(\nu,\nu_\tau))^2 - \frac{1}{2}(\tau|\nabla_W E(\mu)| - W_2(\mu,\mu_\tau))^2 \\
&\quad -\frac{\lambda\tau}{2}\left[2\Lambda_\tau(\mu_\tau,\nu_\tau) + W_2^2(\mu,\nu_\tau) + W_2^2(\nu,\mu_\tau)\right]\ .
\end{aligned}$$

Rearranging terms, we have

$$\begin{aligned}
(1+\lambda\tau)\Lambda_\tau(\mu_\tau,\nu_\tau) &\leq \Lambda_\tau(\mu,\nu) - \frac{1}{2}(\tau|\nabla_W E(\nu)| - W_2(\nu,\nu_\tau))^2 - \frac{1}{2}(\tau|\nabla_W E(\mu)| - W_2(\mu,\mu_\tau))^2 \\
&\quad -\frac{\lambda\tau}{2}\left[W_2^2(\mu,\nu_\tau) + W_2^2(\nu,\mu_\tau)\right]\ .
\end{aligned} \tag{3.4}$$

By the triangle inequality,

$$\begin{aligned} W_2^2(\mu,\nu_\tau) &\geq W_2^2(\mu,\nu) + W_2^2(\nu,\nu_\tau) - 2W_2(\mu,\nu)W_2(\nu,\nu_\tau) \\ &\geq W_2^2(\mu,\nu) - 2W_2(\mu,\nu)W_2(\nu,\nu_\tau) \\ &\geq W_2^2(\mu,\nu) - 2\Lambda_\tau^{1/2}(\mu,\nu)W_2(\nu,\nu_\tau) \ , \end{aligned}$$

and we have a similar bound for $W_2^2(\mu_\tau,\nu)$.

Finally, for $\lambda\tau \leq 1$,

$$\begin{aligned} \frac{1}{2}(\tau|\nabla_W E(\mu)| - W_2(\mu,\mu_\tau))^2 &\geq \lambda\tau\left(\frac{\tau^2}{2}|\nabla_W E(\mu)|^2 - \tau|\nabla_W E(\mu)|W_2(\mu_\tau,\mu)\right) \\ &\geq \lambda\tau\left(\frac{\tau^2}{2}|\nabla_W E(\mu)|^2 - \sqrt{2}\Lambda^{1/2}(\mu,\nu)W_2(\mu_\tau,\mu)\right) \ , \end{aligned}$$

and again we have the same inequality with μ in place of ν. Using these inequalities in (3.4) we obtain the desired bound. □

4 Examples and Applications

4.1 Inequalities (1.16) and (1.22) are Sharp

Our first example shows that the inequality (1.16) from Theorem 1.1 and the inequality (1.22) from Theorem 1.3 are both sharp. For $\lambda \in \mathbb{R}$, consider the functional $E : \mathcal{P}_2^a(\mathbb{R}^d) \to \mathbb{R}$ defined by

$$E(\mu) = \int \frac{\lambda x^2}{2} \mathrm{d}\mu \ . \tag{4.1}$$

As shown in [2, Example 9.3.1], E is proper, coercive, lower semicontinuous, and λ-convex along generalized geodesics.

4.1 PROPOSITION. *For E given by (4.1), $\lambda \geq 0$, and $\tau > 0$, define $\lambda_\tau := \frac{\lambda}{1+\lambda\tau}$. Then E_τ is λ_τ-convex, and no more.*

4.2 PROPOSITION. *For E given by (4.1), $\mu,\nu \in D(E)$, and $\tau > 0$ small enough so that $\lambda\tau > -1$, there is equality in (1.22).*

We first prove the following lemma. For E given by (4.1), it is well-known that the proximal map is simply a scale transformation:

4.3 LEMMA. *For E given by (4.1), $\mu \in D(E)$, and $\tau > 0$ small enough so that $\lambda\tau > -1$, the proximal map associated to E is the scale transformation*

$$\mu \mapsto (1+\lambda\tau)^{-1}\mathrm{id}\#\mu \tag{4.2}$$

where $\mathrm{id}(x) = x$ *is the identity transformation. Moreover, for any $\mu,\nu \in D(E)$,*

$$W_2^2(\mu_\tau,\nu_\tau) = \frac{1}{(1+\lambda\tau)^2} W_2^2(\mu,\nu) \tag{4.3}$$

and

$$W_2^2(\mu,\nu_\tau) = \frac{1}{1+\lambda\tau}\left[W_2^2(\mu,\nu) + 2\tau\left(E(\mu) - \frac{1}{1+\lambda\tau}E(\nu)\right)\right] \ . \tag{4.4}$$

Proof. At any $\mu \in D(E)$,

$$\nabla \frac{\delta E}{\delta \rho}(\mu) = \nabla \frac{\lambda x^2}{2} = \lambda x \ . \tag{4.5}$$

For $\tau > 0$ small enough so that $\lambda\tau > -1$, the Euler-Lagrange equation (2.5) becomes

$$\mathbf{t}^{\mu}_{\mu_\tau}(x) = x + \lambda\tau x = (1+\lambda\tau)x,$$

μ_τ-almost everywhere. This shows (4.2):

$$(1+\lambda\tau)^{-1}\mathrm{id}\#\mu = \mu_\tau \ .$$

Next, fix $\phi : \mathbb{R}^d \to \mathbb{R}$ convex and define $\nu := \nabla\phi\#\mu$. By uniqueness in the Brenier-McCann theorem, $\nabla\phi$ is the optimal transport map from μ to ν. If ψ is defined by

$$\psi(x) = (1+\lambda\tau)^{-2}\phi((1+\lambda\tau)x) \ ,$$

ψ is convex and $\nabla\psi\#\mu_\tau = \nu_\tau$. Again, by uniqueness in the Brenier-McCann Theorem, $\nabla\psi$ is the optimal transport map between μ_τ and ν_τ. Consequently,

$$\begin{aligned} W_2^2(\mu_\tau,\nu_\tau) &= \int_{\mathbb{R}^d} |\nabla\psi(x) - x|^2 \mathrm{d}\mu_\tau \\ &= (1+\lambda\tau)^{-2} \int_{\mathbb{R}^d} |\nabla\phi((1+\lambda\tau)x) - (1+\lambda\tau)x|^2 \mathrm{d}\mu_\tau \\ &= (1+\lambda\tau)^{-2} \int_{\mathbb{R}^d} |\nabla\phi(x) - x|^2 \mathrm{d}\mu \\ &= (1+\lambda\tau)^{-2} W_2^2(\mu,\nu) \ . \end{aligned}$$

This proves (4.3).

Finally, note that if ϕ is convex and $\nabla\phi\#\mu = \nu$, by the definition of $W_2^2(\mu,\nu)$ and of E,

$$2\int_{\mathbb{R}^d} x\cdot\nabla\phi(x)\mathrm{d}\mu = \frac{2}{\lambda}(E(\mu)+E(\nu)) - W_2^2(\mu,\nu) \ . \tag{4.6}$$

Using that

$$(1+\lambda\tau)^{-1}\nabla\phi\#\mu = \nu_\tau \ ,$$

we may argue as above to show

$$\begin{aligned} W_2^2(\mu,\nu_\tau) &= \int_{\mathbb{R}^d} |(1+\lambda\tau)^{-1}\nabla\phi(x) - x|^2 \mathrm{d}\mu \\ &= \frac{2}{\lambda}(1+\lambda\tau)^{-2}E(\nu) + \frac{2}{\lambda}E(\mu) - 2(1+\lambda\tau)^{-1}\int_{\mathbb{R}^d} x\cdot\nabla\phi(x)\mathrm{d}\mu \ . \end{aligned}$$

Combining this with (4.6) proves (4.4). □

Proof of Proposition 4.1: We first explicitly compute the Moreau-Yosida regularization of E. It follows from (4.2) and the definition of E that for all $\mu \in D(E)$ and $0 < \tau < \infty$,

$$W_2^2(\mu,\mu_\tau) = 2\lambda\tau^2 E(\mu_\tau) \ . \tag{4.7}$$

Again by (4.2),

$$E(\mu_\tau) = (1+\lambda\tau)^{-2}E(\mu) \ . \tag{4.8}$$

Hence,

$$E_\tau(\mu) = \frac{1}{2\tau}W_2^2(\mu,\mu_\tau) + E(\mu_\tau) = (1+\lambda\tau)E(\mu_\tau) = \frac{1}{1+\lambda\tau}E(\mu) \ .$$

Thus, the Moreau-Yosida regularization of E in this (already very regular) case simply multiplies E by a constant.

It is a standard result (see e.g. [2]) that E is λ-convex, and no more. (Its Hessian with respect to the W_2 Riemannain metric is λ times the identity.) It then follows immediately from $E_\tau(\mu) = \dfrac{1}{1+\lambda\tau}E(\mu)$ that E_τ is no more than λ_τ-convex. □

Proof of Proposition 4.2: We proceed by using Lemma 4.3 to express quantities appearing on either side of (1.22) in terms of $W_2^2(\mu,\nu)$, $E(\mu)$ and $E(\nu)$. By the symmetry of μ and ν, equations (4.3) and (4.4) allow us to express $W_2^2(\mu_\tau,\nu_\tau)$, $W_2^2(\mu,\nu_\tau)$ and $W_2^2(\nu,\mu_\tau)$ in these terms. By (2.6), (4.5), (4.7), and (4.8),

$$\tau^2|\nabla_W E(\mu)|^2 = \tau^2\int(\lambda x)^2 d\mu = 2\lambda\tau^2 E(\mu) \quad \text{and} \quad \tau^2|\nabla_W E(\mu_\tau)|^2 = W_2^2(\mu,\mu_\tau) = 2\lambda\tau^2 E(\mu)/(1+\lambda\tau)^2 \ .$$

Symmetric identities hold with ν in place of μ.

Finally, direct calculation shows that both sides of (1.22) are equal to

$$-\frac{2\lambda\tau+\lambda^2\tau^2}{(1+\lambda\tau)^2}\left[W_2^2(\mu,\nu) + \lambda\tau^2(E(\mu)+E(\nu))\right] \ .$$

□

As we see from (4.3), the proximal map for E is always contracting in the W_2 metric for $\lambda > 0$. Thus, in this example, the additional terms in Λ_τ are not required to produce contraction. The point of this example is rather to show that (1.16) and (1.22) are sharp.

4.2 The Discrete Gradient Flow for Entropy and Rényi Entropies

In our second example, we consider functionals E_p corresponding to the entropy and Rényi entropies. We apply Theorem 1.3 to obtain a sharp bound, *uniformly* in the steps of the discrete gradient flow sequence, on the rate at which rescaled solutions of the discrete gradient flow converge to certain limiting densities, known as *Barenblatt densities*. This result mirrors a well-known result obtained by Otto for the corresponding continuous gradient flow. In carrying out this analysis, we learn that the discrete gradient flow is surprisingly well-behaved, not only on average, but also uniformly in the steps. We also show that Otto's beautiful sharp results for the continuous gradient flow can be obtained very efficiently from the analysis of the discrete flow.

First, we define the functionals to be considered. For $p > 1 - 1/d$,[9] define $U_p : \mathbb{R}_+ \to \mathbb{R}$ by

$$U_p(s) := \begin{cases} \frac{s^p - s}{p-1} & \text{if } p \neq 1 \\ s\log s & \text{if } p = 1. \end{cases}$$

[9]The borderline case $p = 1 - 1/d$ is more involved, and, for the sake of simplicity, we do not consider it in this paper. It may be possible to extend our approach to this case using the regularization techniques developed in [4].

Let $\mathcal{P}_2^a(\mathbb{R}^d)$ be the set of probability measures with finite second moment that are absolutely continuous with respect to the Lebesgue measure. Define the functional $E_p : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ by

$$E_p(\mu) := \begin{cases} \int_{\mathbb{R}^d} U_p(f(x))dx & \text{if } \mu \in \mathcal{P}_2^a(\mathbb{R}^d), d\mu(x) = f(x)dx \\ \infty & \text{otherwise.} \end{cases}$$

For $p = 1$, E_p is minus the entropy. For $p \neq 1$, E_p is minus the Rényi entropy. As shown in [2, Example 9.3.6], E_p is proper, lower semicontinuous, and convex along generalized geodesics. As for coercivity, for $p > 1$, E_p is bounded below by $-1/(p-1)$, hence coercive. For $1 - 1/d < p < 1$, E_p is not bounded below, since $\int_{\mathbb{R}^d} f^p(x)\mathrm{d}x$ can be arbitrarily large. E_1 is neither bounded above nor below. Nevertheless, E_p is coercive for $p > 1 - 1/d$, when $d \geq 2$, and for $p > 1/3$, when $d = 1$. Later, we shall need some of the estimates that imply this, so we now explain this case. The $p = 1$ case can be found in [10].

By Hölder's inequality, with exponents $1/p$ and $1/(1-p)$, for all $\nu \in \mathcal{P}_2^a(\mathbb{R}^d)$ with $\mathrm{d}\nu = f(x)\mathrm{d}x$,

$$\begin{aligned} \int_{\mathbb{R}^d} f^p(x)\mathrm{d}x &= \int_{\mathbb{R}^d} f^p(x)(1+|x|^2)^p(1+|x|^2)^{-p}\mathrm{d}x \\ &\leq \left(\int_{\mathbb{R}^d} f(x)(1+|x|^2)\mathrm{d}x\right)^p \left(\int_{\mathbb{R}^d} (1+|x|^2)^{-p/(1-p)}\mathrm{d}x\right)^{1-p} . \end{aligned}$$

Furthermore, $\int_{\mathbb{R}^d} f(x)|x|^2\mathrm{d}x = \int_{\mathbb{R}^d} |x|^2\mathrm{d}\nu = W_2^2(\nu, \delta_0)$, where δ_0 is the Dirac mass at the origin. By the triangle inequality, for any $\mu \in \mathcal{P}_2^a(\mathbb{R}^d)$,

$$W_2(\nu, \delta_0) \leq W_2(\mu, \nu) + W_2(\mu, \delta_0) ,$$

so that

$$\int_{\mathbb{R}^d} f^p(x)\mathrm{d}x \leq \left(\int_{\mathbb{R}^d} (1+|x|^2)^{-p/(1-p)}\mathrm{d}x\right)^{1-p} (1 + (W_2(\mu, \nu) + W_2(\mu, \delta_0))^2)^p .$$

Finally, defining

$$C_p := \frac{1}{1-p}\left(\int_{\mathbb{R}^d} (1+|x|^2)^{-p/(1-p)}\mathrm{d}x\right)^{1-p} ,$$

we have for all $\mu, \nu \in \mathcal{P}_2^a(\mathbb{R}^d)$,

$$E_p(\nu) \geq -C_p\left(1 + 2\int_{\mathbb{R}^d} |x|^2\mathrm{d}\mu + 2W_2^2(\mu, \nu)\right)^p . \tag{4.9}$$

Thus, for all $\mu, \nu \in \mathcal{P}_2^a(\mathbb{R}^d)$,

$$\frac{1}{2\tau}W_2^2(\mu, \nu) + E_p(\nu) \geq \frac{1}{2\tau}W_2^2(\mu, \nu) - C_p\left(1 + 2\int_{\mathbb{R}^d} |x|^2\mathrm{d}\mu + 2W_2^2(\mu, \nu)\right)^p . \tag{4.10}$$

For fixed μ, the right hand side is bounded below for all $\tau > 0$ and $\nu \in P_2^a(\mathbb{R}^d)$, hence E_p is coercive.

Note that the condition $p > 1 - 1/d$, when $d \geq 2$, and $p > 1/3$, when $d = 1$, is exactly the condition to ensure C_p is finite, and it is easy to see that coercivity fails when this is not the case. For a more general result, see [2, Remark 9.3.7].

From this analysis, we may also extract an upper bound on $W_2^2(\mu, \mu_\tau)$ which will be useful later.

4.4 LEMMA (Distance bound for the proximal map)**.** *If $d \geq 2$, fix $p > 1 - 1/d$, and if $d = 1$, fix $p > 1/3$. Let $\mu \in D(E_p)$ and*

$$M(\mu) := 1 + 2\int_{\mathbb{R}^d} |x|^2 \mathrm{d}\mu \ .$$

Then for all τ small enough that $4pC_p\tau < 1$,

$$W_2^2(\mu, \mu_\tau) \leq 2\tau \frac{E_p(\mu) + C_p M(\mu)}{1 - 4pC_p\tau} \ .$$

A similar, but more complicated, bound in terms of the same quantities holds for all $\tau > 0$.

Proof. By the definition of the proximal map, taking $\nu = \mu$ in the variational problem (1.7), we obtain

$$E_p(\mu) \geq \frac{1}{2\tau} W_2^2(\mu, \mu_\tau) + E_p(\mu_\tau) \ .$$

Then, by (4.10) with $\nu = \mu_\tau$ and Bernoulli's inequality, $(1+u)^p \leq 1 + pu$,

$$\begin{aligned} E_p(\mu) &\geq \frac{1}{2\tau} W_2^2(\mu, \mu_\tau) - C_p \left(M(\mu) + 2W_2^2(\mu, \mu_\tau) \right)^p \\ &= \frac{1}{2\tau} W_2^2(\mu, \mu_\tau) - C_p M^p(\mu) \left(1 + \frac{2W_2^2(\mu, \mu_\tau)}{M(\mu)} \right)^p \\ &\geq \frac{1}{2\tau} W_2^2(\mu, \mu_\tau) - C_p M^p(\mu) \left(1 + p\frac{2W_2^2(\mu, \mu_\tau)}{M(\mu)} \right) \ . \\ &\geq \left[\frac{1}{2\tau} - 2pC_p \right] W_2^2(\mu, \mu_\tau) - C_p M(\mu) \ . \end{aligned}$$

In the last line, we used that $M(\mu) \geq 1$.

The bound is simple due to the use of Bernoulli's inequality $(1+u)^p \leq 1 + pu$. Avoiding this, one obtains a bound without restriction on τ. Since we are mostly concerned with small τ, we leave the details to the reader.

□

If $d \geq 2$, fix $p > 1 - 1/d$, and if $d = 1$, fix $p > 1/3$. Then, E_p is proper, coercive, lower semicontinuous, and convex along generalized geodesics. Therefore, Theorem 2.1 guarantees that the proximal map and discrete gradient flow (1.10) are well-defined for $0 < \tau < \infty$, $\mu_0 \in \overline{D(E_p)}$. Before turning to the long-time asymptotics of the discrete gradient flow for E_p, we first investigate the contraction properties of $\Lambda_\tau(\mu, \nu)$ under the proximal map.

Unlike the functional considered in Section 4.1, E_p is translation invariant. Specifically, for fixed $x_0 \in \mathbb{R}^d$, if T_{x_0} is the translation given by

$$T_{x_0}\mu := (\mathrm{id} - x_0)\#\mu \ ,$$

then $E_p(T_{x_0}\mu) = E_p(\mu)$. The 2-Wasserstein distance is also translation invariant: for any $\mu, \nu \in \mathcal{P}_2^a(\mathbb{R}^d)$

$$W_2^2(\mu, \nu) = W_2^2(T_{x_0}\mu, T_{x_0}\nu) \ .$$

Consequently, the proximal map associated to E_p commutes with translations:

$$(T_{x_0}\mu)_\tau = T_{x_0}(\mu_\tau) \ .$$

On one hand, this implies that the proximal map does not contract strictly in W_2^2: for any $\nu \in \mathcal{P}_2^a(\mathbb{R}^d)$, $W_2^2(\nu, T_{x_0}\nu) = x_0^2$, so

$$W_2^2(\mu_\tau, (T_{x_0}\mu)_\tau) = W_2^2(\mu, T_{x_0}\mu) \ .$$

On the other hand, because the functional E_p is strictly convex [2, 19], strict inequality holds in (3.3) and hence in (1.21) of Theorem 1.3:

$$\Lambda_\tau(\mu_\tau, \nu_\tau) < \Lambda_\tau(\mu, \nu) \ .$$

Therefore, $\Lambda_\tau(\mu, \nu)$ is strictly decreasing under the proximal map, even though $W_2^2(\mu, \nu)$ is not.

We now turn to the long-time asymptotics of the discrete gradient flow for E_p. As shown by Otto [19], the $\tau \to 0$ limit of the discrete gradient flow tends to the continuous gradient flow on $\mathcal{P}_2^a(\mathbb{R}^d)$, which corresponds to the porous medium equation or the fast diffusion equation:

$$\frac{\partial}{\partial t}\rho(t,x) = \Delta \rho(t,x)^p \ . \tag{4.11}$$

(For $p < 1$ this is the fast diffusion equation. For $p > 1$, it is the porous medium equation.) We show that for each $\tau > 0$, the discrete flow is a strikingly close analogue of the continuous flow.

A key feature of (4.11) is that it has *self-similar scaling solutions* known as *Barenblatt solutions*,

$$\sigma_p(t,x) := t^{-d\beta} h_p\left(\frac{x}{t^\beta}\right) \ , \tag{4.12}$$

where

$$\beta := \frac{1}{2 + d(p-1)} \ , \tag{4.13}$$

and

$$h_p(x) := \begin{cases} (\lambda + \frac{1-p}{p}\frac{\beta}{2}|x|^2)^{1/(p-1)} & \text{if } 1 - \frac{1}{d} < p < 1 \\ \lambda\ e^{-\beta|x|^2/2} & \text{if } p = 1 \\ (\lambda + \frac{1-p}{p}\frac{\beta}{2}|x|^2)_+^{1/(p-1)} & \text{if } p > 1, \end{cases} \tag{4.14}$$

with normalizing constants $\lambda = \lambda(d,p)$ so that $\int_{\mathbb{R}^d} d\sigma_p(x) = \int_{\mathbb{R}^d} h_p(x)dx = 1$.

4.5 DEFINITION (Barenblatt density)**.** If μ is a probability measure of the form $d\mu = \sigma_p(t,x)dx$, we call μ a *Barenblatt density*. Going forward, we will simply write $\mu = \sigma_p(t,x)dx$.

We now show that the Barenblatt densities are preserved under the discrete gradient flow. Before stating the next proposition, let us observe that $0 < \beta < 1$ for all values of $p > 1 - 1/d$. Thus, the function $s \mapsto s^\beta - \tau\beta s^{\beta-1}$ is strictly monotone increasing for $s \geq 0$ and yields the value 0 for $s = \tau\beta$. Consequently, for any $r > 0$, there is a unique $s > \tau\beta$ such that

$$r^\beta = s^\beta - \tau\beta s^{\beta-1} \ . \tag{4.15}$$

4.6 DEFINITION (Proximal time-shift function)**.** Define the proximal time-shift function $\theta_\tau : \mathbb{R}_+ \to \mathbb{R}_+$ so that, for any $r > 0$, $\theta_\tau(r)$ is the unique value of s that solves (4.15).

We have already observed that $\theta_\tau(r) > \tau\beta$ for all $r > 0$. Since $r^\beta - \tau\beta r^{\beta-1} < r^\beta$ for all $r > 0$, $\theta_\tau(r) > r$. The following lemma generalizes a result in [7] for the $p = 1$ case, showing that the proximal map for the functional E_p takes $\sigma_p(r,x)\mathrm{d}x$ to $\sigma_p(\theta_\tau(r),x)\mathrm{d}x$. Thus the proximal map takes a Barenblatt density to a Barenblatt density with a larger "time parameter". Given that the class of Barenblatt densities is preserved at the discrete level, we would of course expect the time parameter to increase.

4.7 PROPOSITION. *If $d \geq 2$, fix $p > 1 - 1/d$, and if $d = 1$, fix $p > 1/3$. Let μ be a Barenblatt density, i.e. $\mu = \sigma_p(r,x)\mathrm{d}x$ for some $r > 0$. Then, for $\tau > 0$, the image of μ under the proximal map for E_p is of the form*

$$\mu_\tau = \sigma_p(\theta_\tau(r),x)\mathrm{d}x \ . \tag{4.16}$$

Proof. Given a Barenblatt density $\mu = \sigma_p(r,x)dx$ for some $r > 0$, let $s := \theta_\tau(r)$ and $\nu := \sigma_p(s,x)\mathrm{d}x$. We compute

$$\nabla\frac{\delta E_p}{\delta\rho}(\nu) = U_p''(\sigma_p(s,x))\nabla\sigma_p(s,x) = p\sigma_p(s,x)^{p-2}\nabla\sigma_p(s,x)(x) = -\frac{\beta x}{s} \quad \nu\text{-almost everywhere,} \tag{4.17}$$

Next, note that since $s = \theta_\tau(r) > \tau\beta$,

$$\nabla\varphi(x) := x + \tau\nabla\frac{\delta E_p}{\delta\rho}(\nu) = \left(1 - \frac{\tau\beta}{s}\right)x$$

is the gradient of a convex function. Consequently, if we define

$$\rho := \nabla\varphi\#\nu \ , \tag{4.18}$$

uniqueness in the Brenier-McCann Theorem guarantees that $\nabla\varphi$ is the optimal transport map between ν and ρ. Since $\nabla\varphi = \mathbf{t}_\nu^\rho = \mathrm{id} + \tau\nabla\frac{\delta E_p}{\delta\rho}(\nu)$ is the the Euler-Lagrange equation (2.5), $\nu = \rho_\tau$, the image of ρ under the proximal map. With the explicit form of $\nabla\varphi$ and $\sigma_p(s,x)$, we compute

$$\rho = \left(1 - \frac{\tau\beta}{s}\right)^{-d}\sigma_p\left(s, \left(1 - \frac{\tau\beta}{s}\right)^{-1}x\right)\mathrm{d}x = \sigma_p\left(\left(1 - \frac{\tau\beta}{s}\right)^{1/\beta}s, x\right)\mathrm{d}x \ .$$

By definition of $s = \theta_\tau(r)$

$$r = \left(1 - \frac{\tau\beta}{s}\right)^{1/\beta}s \ .$$

Therefore, $\rho = \sigma_p(r,x)\mathrm{d}x = \mu$, so $\mu_\tau = \rho_\tau = \nu = \sigma_p(s,x)\mathrm{d}x = \sigma_p(\theta_\tau(r),x)\mathrm{d}x$. □

Note that when τ is very small compared to $t > 0$, and hence also compared to $s := \theta_\tau(t)$,

$$t = \left(1 - \frac{\tau\beta}{s}\right)^{1/\beta}s \approx s - \frac{\tau\beta}{\beta} = s - \tau \ ,$$

so $\theta_\tau(t) \approx t + \tau$. Thus, in this approximation, the proximal map shifts the time forward by τ, independent of t. To the extent this is accurate, it makes it very easy to understand the discrete gradient flow for E_p starting from a Barenblatt density: at the nth step of size τ, one gets a Barenblatt density whose time parameter has been increased by approximately $n\tau$. The following lemma allows us to control this approximation in precise terms.

4.8 LEMMA. *Fix $r > 0$. Then, for all $t \geq r$,*

$$\left(\frac{r}{r+\tau}\right)\tau \;\leq\; \theta_\tau(t) - t \;\leq\; \tau \tag{4.19}$$

Proof. Let $s := \theta_\tau(t)$ for any $t \geq r$. We recall that $0 < \beta < 1$ for all $p > 1 - 1/d$. By the definition of θ_τ, we have

$$t^\beta = s^\beta - \tau\beta s^{\beta-1} \ .$$

Assume $s > t + \tau$. Then, by Bernoulli's inequality $(1+u)^{1-\beta} \leq (1 + (1-\beta)u)$ with $u := \tau/t$,

$$t^\beta = s^\beta - \tau\beta s^{\beta-1} > (t+\tau)^\beta - \tau\beta(t+\tau)^{\beta-1} = (t+\tau)^{\beta-1}(t + (1-\beta)\tau) = t^\beta(1+u)^{\beta-1}(1+(1-\beta)u) \geq t^\beta \ .$$

This is a contradiction. Therefore, $\theta_\tau(t) = s \leq t + \tau$, which proves the upper bound in (4.19).

To obtain the lower bound, we use the upper bound on s and the relation $s = t\,(1 - \tau\beta/s)^{-1/\beta}$ to obtain $s \geq t\left(1 - \dfrac{\tau\beta}{t+\tau}\right)^{-1/\beta}$. Then since $(1+u)^{-1/\beta} \geq 1 - u/\beta$ and $t \geq r$,

$$s \geq t\left(1 + \frac{1}{\beta}\frac{\tau\beta}{t+\tau}\right) \geq t + \tau\left(\frac{r}{r+\tau}\right) \ .$$

□

We may now use Theorem 1.3 to control the rate at which rescaled solutions to the discrete gradient flow converge to a Barenblatt density. First, we define the rescaled discrete gradient flow. For any positive integer n, let θ_τ^n be the n-fold power of θ_τ. For $t > 0$, let S_t denote the scaling transformation given by

$$S_t\nu = \frac{\mathrm{id}}{t^\beta}\#\nu \ .$$

Since $t^{-\beta}x$ is the gradient of a convex function, uniqueness in the Brenier-McCann Theorem implies that it is the optimal transport map from ν to $S_t\nu$.

Let μ be a Barenblatt density, i.e., $\mu = \sigma_p(r,x)\mathrm{d}x$ for some $r > 0$. Then $S_r\mu = h_p(x)dx$. Let $\{\mu_n\}$ be the discrete gradient flow with initial data μ for fixed $\tau > 0$. By Proposition 4.7,

$$J_\tau^n\mu = \mu_n = \sigma_p(\theta_\tau^n(r), x)\mathrm{d}x \ ,$$

and by definition of the scaling transformation,

$$S_{\theta_\tau^n(r)}J_\tau^n\mu = S_{\theta_\tau^n(r)}\mu_n = h_p(x)\mathrm{d}x \qquad \text{for all } n \in \mathbb{N} \ . \tag{4.20}$$

Thus, each step of the discrete gradient flow sequence is also a rescaling of $h_p(x)\mathrm{d}x$.

In fact, something almost as good holds even when the initial data of the discrete gradient flow is not a Barenblatt density. We apply Theorem 1.3 to prove that if $\{\nu_n\}$ is a discrete gradient flow with initial data $\nu \in D(E_p)$ for fixed $\tau > 0$, then

$$\lim_{n\to\infty} S_{\theta_\tau^n(r)}J_\tau^n\nu = \lim_{n\to\infty} S_{\theta_\tau^n(r)}\nu_n = h_p(x)\mathrm{d}x \ .$$

That is, if you wait a while and scale the solution to view it in a fixed length scale, what you see is (essentially) a Barenblatt density, no matter what the initial data $\nu \in D(E_p)$ looked like. Moreover, we show that $W_2(S_{\theta_\tau^n(r)}\nu_n, h_p(x)\mathrm{d}x)$ essentially contracts at a precise polynomial rate.

4.9 THEOREM (Discrete fast diffusion and porous medium flow)**.** *If $d \geq 2$, fix $p > 1 - 1/d$, and if $d = 1$, fix $p > 1/3$. Let $\nu \in D(E_p)$ and let $\mu = \sigma_p(r, x)dx$ for some $r > 0$. Given $0 < \tau \leq 1$, let $\{\nu_n\}$ and $\{\mu_n\}$ be the discrete gradient flows (1.10) with initial conditions ν and μ. Define the rescaled discrete gradient flow sequence*

$$\widetilde{\nu}_n := S_{\theta_\tau^n(r)}\nu_n \ .$$

Then, there is an explicitly computable constant K depending only on d, p, r, $E_p(\nu)$, and

$$M(\nu) := 1 + 2\int_{\mathbb{R}^d} |x|^2 \mathrm{d}\nu \ ,$$

so that

$$W_2^2(\widetilde{\nu}_n, h_p(x)\mathrm{d}x) \leq (\theta_\tau^n(r))^{-2\beta}[W_2(\nu,\mu)[W_2(\nu,\mu) + \tau^{1/2}K] + \tau K] \ . \tag{4.21}$$

From this, we readily recover Otto's contraction result for a continuous gradient flow as follows. For any $t > 0$, let $\mathrm{int}(t/\tau)$ denote the integer part of t/τ. By Lemma 4.8, $\theta_\tau(t) = t + \tau$, up to an error that vanishes uniformly in t as $\tau \to 0$. Thus, a simple iteration yields

$$\lim_{\tau \downarrow 0} \theta_\tau^{\mathrm{int}(t/\tau)}(r) = r + t \ . \tag{4.22}$$

Interpolating and taking the limit $\tau \to 0$ as in [10], one obtains from $\{\nu_n\}$ a solution $\rho(t, x)$ to $\dfrac{\partial}{\partial t}\rho(t, x) = \Delta\rho(t, x)^p$ with $\rho(0, x)\mathrm{d}x = \nu_0$. Define the rescaled solution

$$\widetilde{\rho}(t, x) := (r + t)^{d\beta}\rho(t, (r + t)^\beta x) \ .$$

We then conclude that, for all $t > 0$,

$$W_2^2(\widetilde{\rho}(t, x)\mathrm{d}x \ , \ h_p(x)\mathrm{d}x) \leq (r + t)^{-2\beta} W_2^2(\rho(0, x)\mathrm{d}x \ , \ \sigma_p(r, x)\mathrm{d}x) \ .$$

One may choose r to minimize $W_2^2(\rho(0, x)\mathrm{d}x \ , \ \sigma_p(r, x)\mathrm{d}x)$. Otto has shown this contraction result is sharp. Hence the "near contraction" result we obtain in the discrete setting cannot be improved in any manner that is uniform in τ.

Other aspects of Otto's analysis that leverage this contraction into a bound on L^1 convergence may be applied at the discrete level without difficulty, and we do not go into the details here. On the other hand, while Otto proves a continuous gradient flow analogue of Theorem 1.3, his proof does not extend to the discrete case. Theorem 1.3 provides the means to carry out the discrete analysis and to show that the discrete gradient flow analogue of (4.11) is surprisingly complete.

Proof of Theorem 4.9. By Theorem 1.3, applied iteratively, we have

$$\Lambda_\tau(\nu_n, \mu_n) \leq \Lambda_\tau(\nu_1, \mu_1) = \Lambda_\tau(\nu_\tau, \mu_\tau) \ . \tag{4.23}$$

Note that we make the comparison with $\Lambda_\tau(\nu_\tau, \mu_\tau)$, not $\Lambda_\tau(\nu, \mu)$, since $|\nabla_W E_p(\nu)|^2$ (and hence $\Lambda_\tau(\mu, \nu)$) may be infinite, but by [2, Theorem 3.1.6], the strict convexity of E implies

$$|\nabla_W E(\nu_\tau)|^2 < |\nabla_W E(\nu)|^2 \tag{4.24}$$

so $\Lambda_\tau(\nu_\tau,\mu_\tau)<\infty$. We shall show that $\Lambda_\tau(\nu_\tau,\mu_\tau)$ is very close to $W_2^2(\nu,\mu)$, differing by a term that is $\mathcal{O}(\tau^{1/2})$. Specifically, there exists a constant K depending only d, p, r, $E_p(\nu)$, and $M(\nu)$, such that

$$\Lambda_\tau(\nu_\tau,\mu_\tau)\le W_2(\nu,\mu)[W_2(\nu,\mu)+\tau^{1/2}K]+\tau K\ . \tag{4.25}$$

Using this in (4.23), we obtain

$$W_2^2(\nu_n,\mu_n)\le\Lambda_\tau(\nu_n,\mu_n)\le W_2(\nu,\mu)[W_2(\nu,\mu)+\tau^{1/2}K]+\tau K\ . \tag{4.26}$$

Next, by the scaling properties of the 2-Wasserstein metric and (4.20), for all $n\ge 1$,

$$(\theta_\tau^n(r))^{-2\beta}W_2^2(\nu_n,\mu_n)=W_2^2(S_{\theta_\tau^n(r)}\nu_n,S_{\theta_\tau^n(r)}\mu_n)=W_2^2(\widetilde{\nu}_n,h_p(x)\mathrm{d}x)\ .$$

Therefore,

$$W_2^2(\widetilde{\nu}_n,h_p(x)\mathrm{d}x)\le(\theta_\tau^n(r))^{-2\beta}[W_2(\nu,\mu)[W_2(\nu,\mu)+\tau^{1/2}K]+\tau K]\ ,$$

which is (4.21).

It remains to prove (4.25). First, note that since $\mu=\sigma_p(r,x)\mathrm{d}x$, (4.17) implies $\nabla\dfrac{\delta E_p}{\delta\rho}(\mu)=-\dfrac{\beta x}{r}$. Thus, by Lemma 2.2 and the definition of the length of the gradient (1.9),

$$\tau^2\frac{\beta^2}{r^2}\int_{\mathbb{R}^d}|x|^2\sigma_p(r,x)\mathrm{d}x=\tau^2|\nabla_W E_p(\mu_\tau)|^2=W_2^2(\mu,\mu_\tau)\ . \tag{4.27}$$

We will consider the cases $p<1$, $p=1$, and $p>1$ separately. For $1-\frac{1}{d}<p<1$, when $d\ge 2$, and $1/3<p<1$, when $d=1$, we may use the bound on $W_2(\nu,\nu_\tau)$ provided by Lemma 4.4 to show

$$\tau^2|\nabla_W E_p(\nu_\tau)|^2\le W_2^2(\nu,\nu_\tau)\le 2\tau\frac{E_p(\nu)+C_pM(\nu)}{1-4pC_p\tau}\ . \tag{4.28}$$

(This particular bound requires $4pC_p\tau<1$, but one may prove a similar bound with a more complicated constant that holds for all $\tau>0$.) By the triangle inequality,

$$\begin{aligned}W_2^2(\mu_\tau,\nu_\tau)&\le(W_2(\mu,\nu)+W_2(\mu,\mu_\tau)+W_2(\nu,\nu_\tau))^2\\&\le W_2^2(\mu,\nu)+2W_2(\mu,\nu)[W_2(\mu,\mu_\tau)+W_2(\nu,\nu_\tau)]+2W_2^2(\mu,\mu_\tau)+2W_2^2(\nu,\nu_\tau)\ .\end{aligned}$$

Combining this with (4.28) and (4.27) gives

$$\begin{aligned}\Lambda_\tau(\mu_\tau,\nu_\tau)&\le W_2^2(\mu,\nu)+2W_2(\mu,\nu)\left[\left(2\tau\frac{E_p(\nu)+C_pM(\nu)}{1-4\tau pC_p}\right)^{1/2}+\tau\frac{\beta}{r}\left(\int_{\mathbb{R}^d}|x|^2\sigma_p(r,x)\mathrm{d}x\right)^{1/2}\right]\\&+5\tau\frac{E_p(\nu)+C_pM(\nu)}{1-4\tau pC_p}+\frac{5}{2}\tau^2\frac{\beta^2}{r^2}\int_{\mathbb{R}^d}|x|^2\sigma_p(r,x)\mathrm{d}x\ .\end{aligned}$$

This leads directly to (4.25) with an explicit constant.

For $p>1$, by Lemma 2.2 and the definition of the proximal map,

$$\tau^2|\nabla_W E_p(\nu_\tau)|^2\le W_2^2(\nu,\nu_\tau)\le 2\tau[E_p(\nu)-E_p(\nu_\tau)]\ .$$

Since E_p is bounded below, an analogous argument leads to (4.25).

The case $p=1$ is similar to the case $p<1$; we leave the details to the reader.

□

Acknowledgement We thank Luigi Ambrosio for helpful comments on a draft of this paper. We thank Haim Brezis for an enlightening conversation. We thank the anonymous reviewers for many useful suggestions.

References

[1] L. Ambrosio and W. Gangbo, Hamiltonian ODEs in the Wasserstein space of probability measures, Comm. Pure Appl. Math. **61** (2008), no. 1, 18–53. MR2361303 (2009b:37101)

[2] L. Ambrosio, N. Gigli and G. Savaré, *Gradient flows in metric spaces and in the space of probability measures*, second edition, Lectures in Mathematics ETH Zürich, Birkhäuser, Basel, 2008. MR2401600 (2009h:49002)

[3] C. Baiocchi, Discretization of evolution variational inequalities, in *Partial differential equations and the calculus of variations, Vol. I*, 59–92, Progr. Nonlinear Differential Equations Appl., 1 Birkhäuser, Boston, Boston, MA. MR1034002 (90m:49002)

[4] A. Blanchet, E. A. Carlen and J. A. Carrillo, Functional inequalities, thick tails and asymptotics for the critical mass Patlak-Keller-Segel model, J. Funct. Anal. **262** (2012), no. 5, 2142–2230. MR2876403

[5] H. Brézis, Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations, in *Contributions to nonlinear functional analysis (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1971)*, 101–156, Academic Press, New York. MR0394323 (52 #15126)

[6] H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland, Amsterdam, 1973. MR0348562 (50 #1060)

[7] E. A. Carlen and W. Gangbo, Constrained steepest descent in the 2-Wasserstein metric, Ann. of Math. (2) **157** (2003), no. 3, 807–846. MR1983782 (2004c:49027)

[8] E. De Giorgi, New problems on minimizing movements, in *Boundary value problems for partial differential equations and applications*, 81–98, RMA Res. Notes Appl. Math., 29 Masson, Paris. MR1260440 (95a:35057)

[9] E. Emmrich, Discrete Versions of Gronwall's Lemma and Their Application to the Numerical Analysis of Parabolic Problems, preprint no. 637 (1999).

[10] R. Jordan, D. Kinderlehrer and F. Otto, The variational formulation of the Fokker-Planck equation, SIAM J. Math. Anal. **29** (1998), no. 1, 1–17. MR1617171 (2000b:35258)

[11] H. K. Kim, *Hamiltonian systems and the calculus of differential forms on the Wasserstein space*, ProQuest LLC, Ann Arbor, MI, 2009. MR2713745

[12] B. Martinet, Régularisation d'inéquations variationnelles par approximations successives, Rev. Française Informat. Recherche Opérationnelle **4** (1970), Ser. R-3, 154–158. MR0298899 (45 #7948)

[13] B. Martinet, Algorithmes pour la résolution des problémes d'optimisation et de minmax, Thése d'état Université de Grenoble (1972).

[14] R. J. McCann, A convexity principle for interacting gases, Adv. Math. **128** (1997), no. 1, 153–179. MR1451422 (98e:82003)

[15] R. J. McCann, Existence and uniqueness of monotone measure-preserving maps, Duke Math. J. **80** (1995), no. 2, 309–323. MR1369395 (97d:49045)

[16] J. J. Moreau, Proximité et dualité dans un espace hilbertien, Bull. Soc. Math. France **93** (1965), 273–299. MR0201952 (34 #1829)

[17] J. J. Moreau, Fonctionnelles convexes, Séminaire sur les Équations aux Dérivées Partielles (19661967). II. (French) Collége de France, Paris (1967).

[18] F. Otto, Doubly degenerate diffusion equations as steepest descent, Manuscript (1996).

[19] F. Otto, The geometry of dissipative evolution equations: the porous medium equation, Comm. Partial Differential Equations **26** (2001), no. 1-2, 101–174. MR1842429 (2002j:35180)

[20] R. T. Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28 Princeton Univ. Press, Princeton, NJ, 1970. MR0274683 (43 #445)

[21] T. Strömberg, The operation of infimal convolution, Dissertationes Math. (Rozprawy Mat.) **352** (1996), 58 pp. MR1387951 (97c:49018)

[22] C. Villani, *Topics in optimal transportation*, Graduate Studies in Mathematics, 58, Amer. Math. Soc., Providence, RI, 2003. MR1964483 (2004e:90003)